Soutenance de thèse
Le 13 décembre 2019

Développement d'outils biostatistiques et bioinformatiques de prédiction et d'analyse des défauts de l'épissage : application aux gènes de prédisposition aux cancers du sein et de l'ovaire

Raphaël Leman

Laboratoire de biologie et génétique du cancer

Inserm U1245 Genomics and Personalized Medecine in Cancer and Neurological Disorders

Centre François Baclesse

Normandie Univ, UNICAEN, Caen

Encadré par : Dr Sophie Krieger

Co-Encadré par : Dr Alexandra Martins

PLAN

INTRODUCTION

- ✓ Épissage
- ✓ Prédiction et analyse
- ✓ Prédisposition aux cancers du sein et de l'ovaire

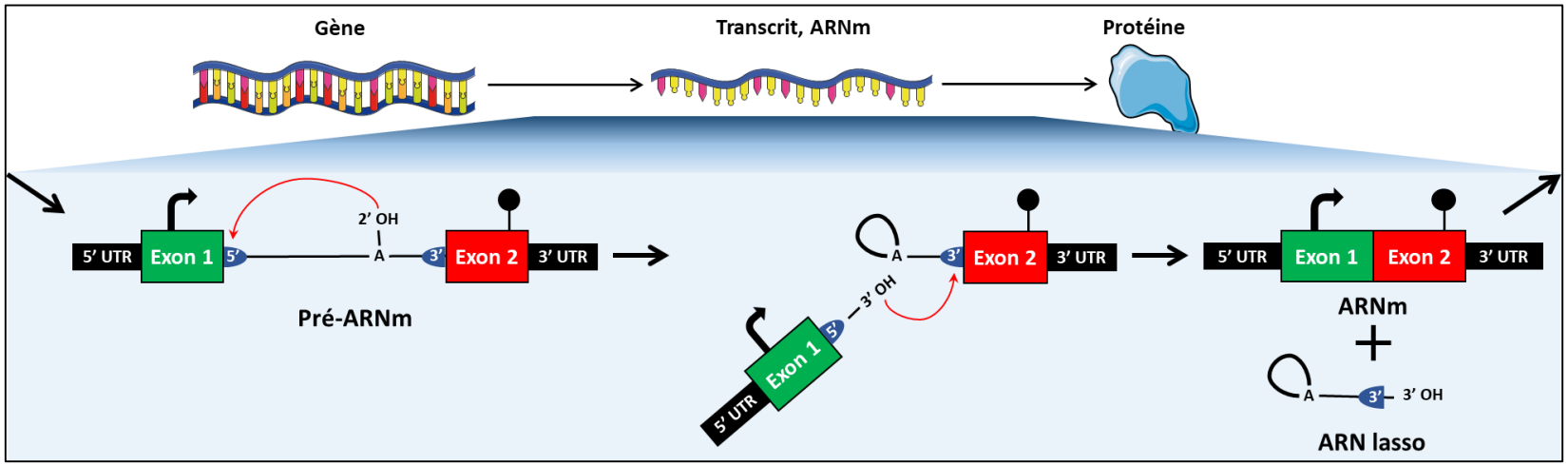
RESULTATS

- ✓ Prédiction des défauts d'épissage
- ✓ Pipeline d'analyse RNA-seq
- ✓ Épissage et règles d'interprétation

DISCUSSION

- ✓ Projet en cours
- ✓ Conclusion

Les motifs d'épissage



Séquence consensus donneur, 5'
Motif hautement conservé (-3/+6)

Point de branchement
Court motif entre -44 et -18

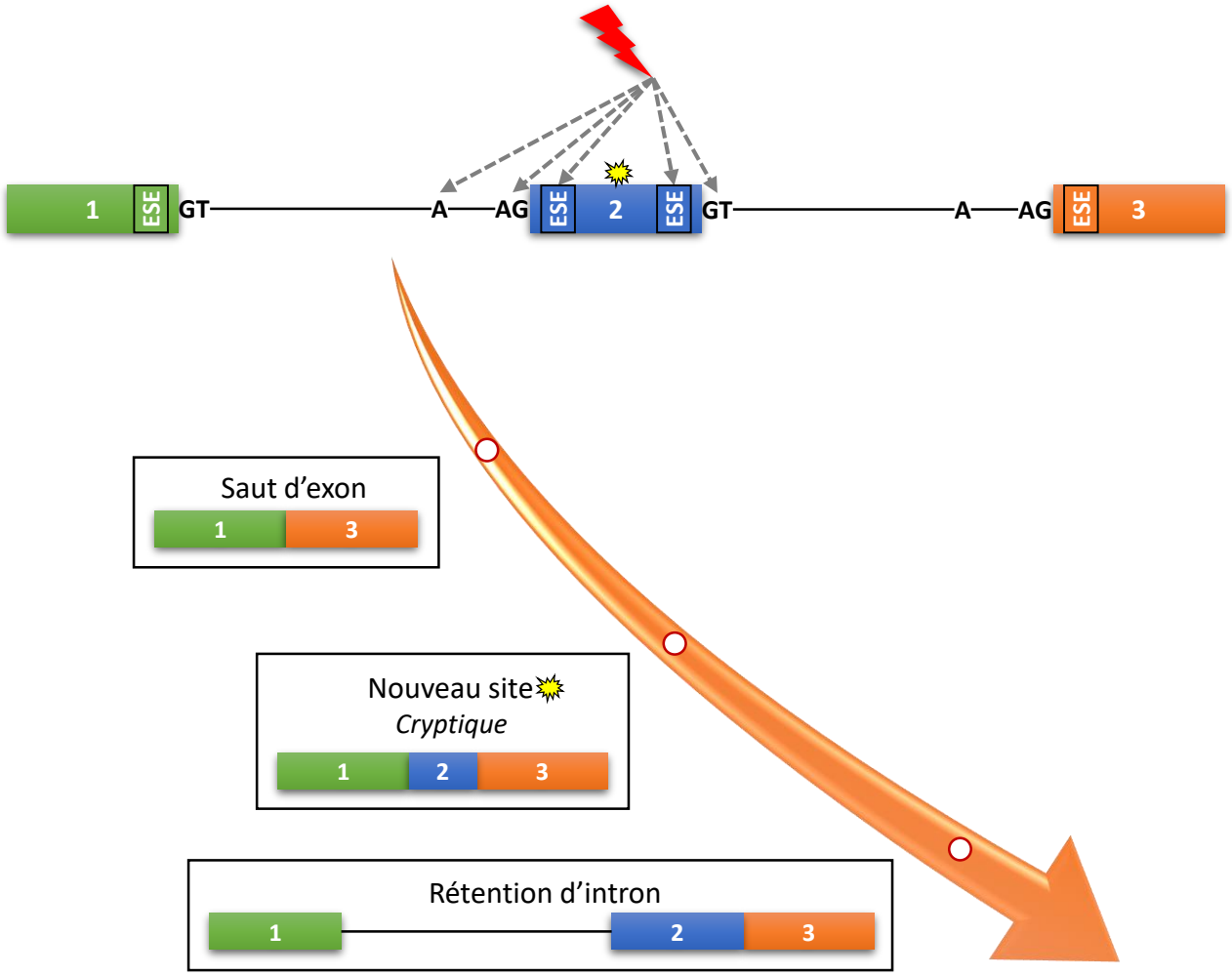
Séquence consensus accepteur, 3'
Motif hautement conservé (-12/+2)

Exonic Splicing Regulators (ESRs): Enhancers/Silencers

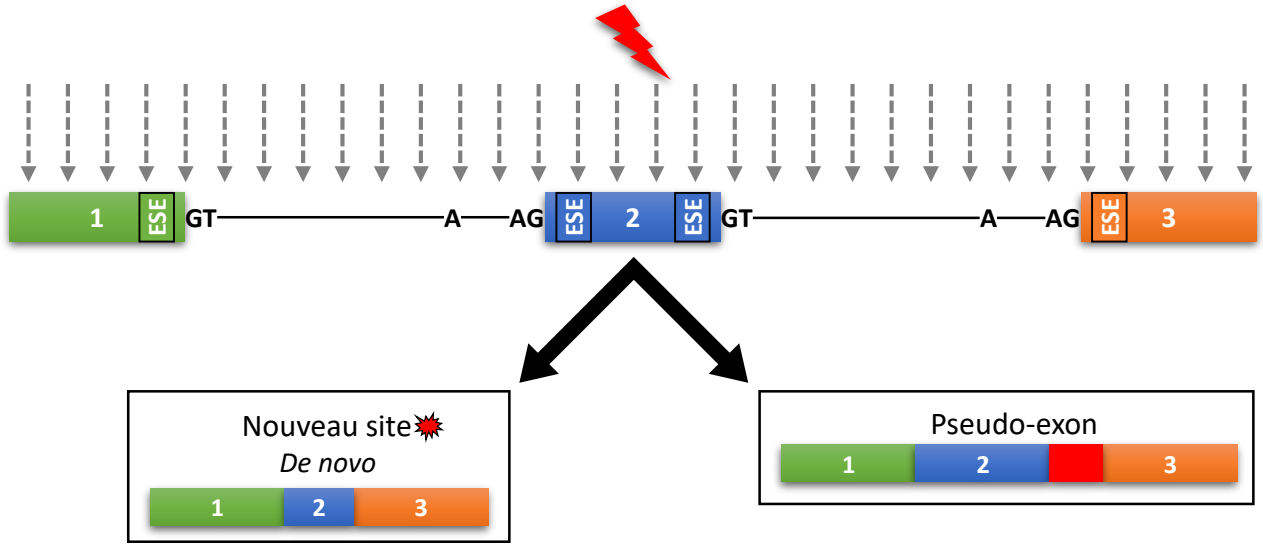
Burge *et al., The RNA World II,* 1999; Mercer *et al., Genome Res.,* 2015; Fairbrother *et al., Science,* 2002

Mécanismes d'altération de l'épissage

Altération des motifs d'épissage



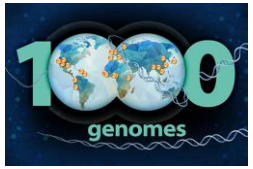
Création de motifs d'épissage



3,8 % des variations génétiques exoniques et introniques proches des exons **impactent l'épissage** en dehors de tout contexte de prédisposition

Cheung *et al., Molecular Cell*, 2018

Comment identifier un variant splicéogénique ?



1000 Genomes Project
➔ 2 500 genomes

Population mondiale
➔ 7 milliards

84 000 000 variations génétiques

$2{,}4 \times 10^{14}$ variations génétiques ⇔ ~1 000 x

Voie lactée
➔ 250 milliards d'étoiles

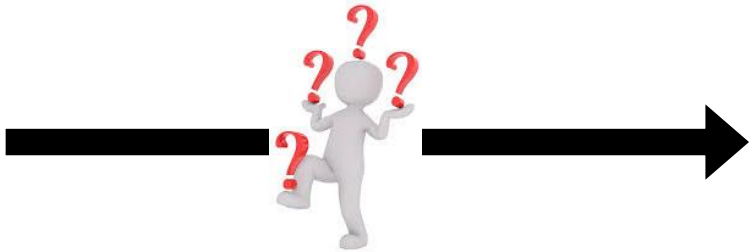
Impossible de tester expérimentalement chaque variant sur l'épissage

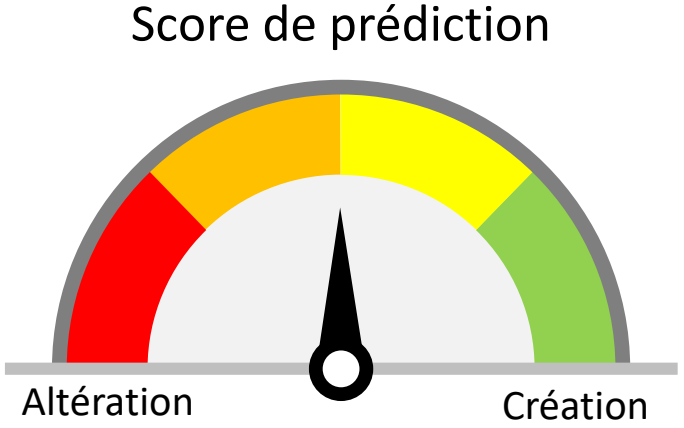
Prédire un variant splicéogénique

La prédiction d'un variant splicéogénique ⇔ la prédiction de la création/altération de motifs d'épissage

TTATGCAGTGGGGATCTCCGGATGCGTAGATTGGATTACACGCAAAAGGACGCTGAGAGTTGACTTTTGAACAATTCTGTGAGGGT
CGTCTAGTACTCCATTTCTATACACCCACTCGTCTGCTAGAACGCGGAGCCGCACGTGTTCTTTTCTACCATGCATGCCTATGTGGAAA
TATCATGTGTCGCTTTGCGGGTTTTTCCTAACCATTTGAGAAGTTTACTCTTCACTTTAGGCAATGTAAGCTGTGGAATCAGCTAGCCT
GCACACCTTGGAACAGAGTCTGTCTATTAAGAGTACACCGACACCCTACCTTGGAAGGACTTGGTCCCCTAATTAATTCATCAATCTT
TAGATGCCTATGAGGGAGTGCATGGGCACGTGACAACGCGAGGTGCTGACCATAGCTGCGAGATTAGGGATAAATGTAGGCCATC
CGTTTAAGCCGGAACTAGCGTAGGCGTGACGGCATGCGAGCGTATGGCACCACACAGGTTACCCCCGACTACGACTGGCACAGCTT
TGTCGACGTTTATTTCTGGGGATGTGGACACCCCTAAATCTGTGAAACGCGCACTACTAACTGAACGCTAAAAGTAAGAGCTCGAGC
TTATCTGCTTTAGCACGTAATTGGTTTTGACTTTCCCAGCATTGGTGCGCACTCGCCGAAATACAATATTAAAGAGTAAGAAAAGGCG
AATCACCAAAGCAACCCCATCCTATACTCTGGAGCTTTAGCAGTAAGGCAATAGGTCCAGTACTCCTGACAAATGGAGAGAAAGGC
GGTCTATGCTCCTCTTTAAACAATGTTTTAGATTAACCTCGGACAACAGAACTTTGGCTCTCCCGATATAGCCAGAATCGTGATGGCC
TGAATAATATATCGTTGCCGCGTTGGGAACTATGCCTGAAGTTACTCTTCGACCGAACCATTTTTAGTGATATCCTGTGGTATAGTCT
AGTTATAGGTCAGGGGCTAGGGTCCAAGTTTCGAGATCAGCAGGTCGCACGATAGATTATACGTAGTCTGAATGTGGGCTTTCGGG
TACGTCTTATGAGGATCTGTCTCACGCCATCGTATTGGGTAAAGGCATGGTCCGCGGGTGGCGATGGAGATGAGGGTGTTAAAATT
ACTGCATGAACGTTGGCCCGCTCCACTGAAGGTATGTGTCCTAGTAGGGATTAAGTGCGGAGACACCCGTACTCGATTCCCAGACCC
TTCCAAATTTGATGCCATTGCACGTGCCTATGGTAGACTAGCTCGCATCGAACAGCCTCTTGTAACTCATGCGGTTGGAGTTCAACGC
TTCCGAGTGAGACGCGTGAATACGGATGGGATCCCTCCTGGGCAGACATAAAAGAATACGTATAATGGGGCTCGTTCCACTTTTTAG
ACGGTCGATTTTCGAGGGGCCGGCCCAACGCTGTAGTGAAAGCGTTGCTGGTCCGATGTACGCCGCTGGGCCAAGGAGCGCGACC
ACTTTTGCCAGGTCAAGTAAGGTGAGCGTACCAGCGCCACGACTGCACTAATCATCGGCTGTATCATAGTCCTTTAGTTCCAGACAG

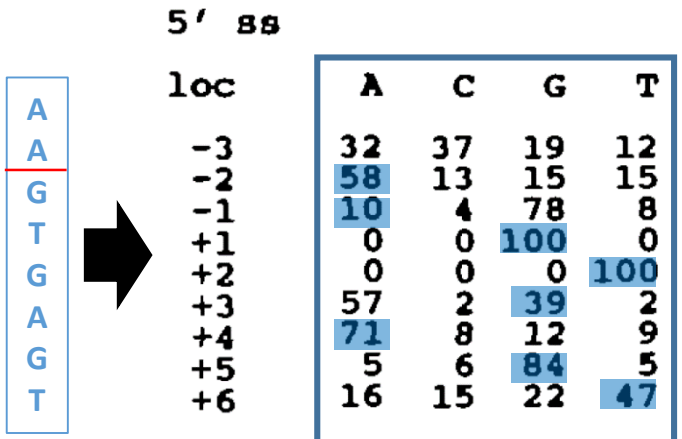
GATACAGTTAGATACGGCACGTTCGGCAATCGCGTTT**G**CGTACGGTTGGTTTCCTTGGCCTCACTATTCACCCGGCGCTGATTGATT
CGGCGTATGGCTATGACTAGGGAGGGTCAATAGGTCGTCTGCACGCCAAGAGCGACGAGAACTATTGAAGAGCCATACAGTAGTCC
TGGCTGCAGGTTGCACATTCGTTGCGATAATAACGTATGAGCGGACCGCGATGGGAGGGTAAAAACCCTTCAATCGAGGGTATGTT
CACCGTTACATACGGCTCCTGTACTTGATCCCGCGTTTCCTGAGCGCTTGGTCAACATAAATATCAATCCCGGACCGAACATGGTAAT
ACTTCAGGCCGTCCGGGGAGAAGTCTCGCATAACCTGAATAACGCGCCAGTAATGGTGAACTCCCCGGTGTTCCATGGCACCCTATG
GTATGAAATCTACTCTGTAAATTGTGTAGCTTGCAAGAGTGGCATTCACATTTTGGCTGCATTTGTATTATTGTCTTACTAACATTGTT
CTATTTCTCTTGTCCGTCTGTAGTGATTTTCCGCTCAGGATATCTACAGTGGGAATAGCAGCGTTGTGTCGATATGGTCGCTTATTGTC
TCGAATTTGTGACTGAGGAGGCTGAGTTGATCCGCGCTGCTTAGCATAAGATGGATAATGGTTAGCTGAGTCAGAGCTGACGGATT
CTCCCCGCCTTTCTGGGTGCAGGTTATCTTCTTCGTTAGATCTAGTCCGCTCATTCGCGTTCGGATCACTAATGTGCTCCCTCCTTCGC
CGTCAGATCGTCGGATGAACGATAGATTGCGACCCAATGTTGGAAGTCGGGCATTCGAGATGGCTCTATCTGGTCTACTGGACTGCT
GTGCAGCCTGAGGTCTACGGAAAAGTTCCCAAGCTTTTGCTCGTCAAGGTTCCCAACAATTGCGCCTCCGTATGGGTTCCCCTCAGA
AGTTCTGTGCCAACCGCGACAGCCTATCTACGCTTGGGCTAATTCAGTGACACGCGCTAGACTGGATTACCTAATCGACCCGGGAGA
GCTAACGACCCAGATCGCAATCGTTTGAGTTCTTACCGACCATCGATGCTGCAAAGAGAACTTCATTTGTGGACTACCTACCAAAGCT
GGGGGCACTCTAGCTGTAAGTGCTTTTAACCCTACGTTAGCTCCATTCCCGCGCTCGATGTGCATCGCCTCACACAGCAAGATTTACT
AAGCCCCTGGTGGCCTCCCTTGGGTTTTGGTTCTAGCCCTAGAAAAGACATCTCTTCTAACAAGCTCATCGTTAGTGACCAGTTTTGT
AGGCTCGGGCTTGACCATTTAATTACCCCTCGGATTTGGGACCTAGAAGAAGGTCGGGAGACCAATAGCAGTTAATTTATCCGAACT
CCTCAGATGTGGTCGCTCACCTAAGGGAAATATTGGAGTTGTAACCGGGGGGTGGCTATTGCCGGTGTGTTGACTACGGACCATGT
GAGACCTGTCTTGCTAGTAGCACACCGCTACCATCTGATCGCTACTGGATTGAGATTGGACCCAATGGCTATAAGGTCTTGCTCAGC
ACATGTCTTAACTCGGAATTAGAGGACGCGTCGAACTAAAACTGAGGTCTTACACCTTACACGCTTTTAGTTTGCAATTCAGCCCGAA
ATTTTGAAAGAAGTGAAGGTTAATGTTCTGTTAAAGAATTAGATATAGCGTGTTCATTAGCAGATCGACAAGCGGTCGGAACGGGA
CCTCATAAAGCGTTGTGATGTTGGTATCGCCGAGGTTTGTATCGGGCACCGTCACTAAAAGATAGATTAGTAACGGCCGTCGTAGAG
TGCGAAATATTCGTAAATCTGAGGGTTACCATACGAAGTGCCGCATCGTGTACTTGACTTCTGGATCCCGAAGTGGATCTTAATACTC
TAGTATGAACAGTTACAGGTGACGGTGCTGCTCCCAAATTCTTTAGACCAATCTCTAGCCTGGTGCCGTGGAACTGCCTTGTCGAGG
TGGTGGGCCCATTACACTCGGATTGAGTTGTTCGTGCCGCCAGAGACCGGTCCGATTCCTCGGTCGCGCGCCCCGCAACG





Un score par motif
Ex: Splice Site Finder (SSF)[1] et MaxEntScan (MES)[2]

SSF



Somme des % minimaux

Somme des % maximaux

$$Score_{SSF} = 100 \times \left(\frac{t - mint}{maxt - mint} \right)$$

$$Score_{SSF} = 84,3\ \%$$

MES

Non dépendante

AGCTGATCAGTCTGATTCC

$P(X_1)$ non dépendant de $P(X_2)$, ….

De proche en proche

AGCTGATCAGTCTGATTCC

$P_{X1}(X_2) > 0$, $P_{X2}(X_3) > 0$, ….

Discontinue

AGCTGATCAGTCTGATTCC

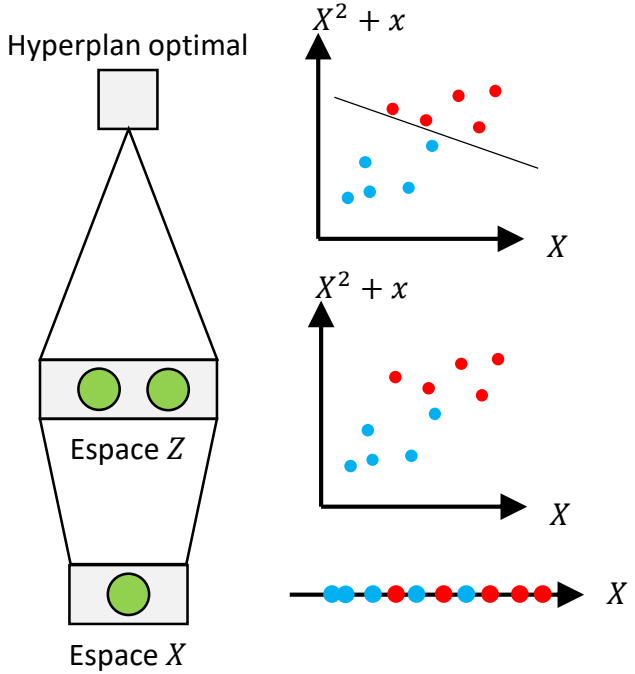
$P_{X2}(X_4) > 0$, $P_{X5}(X_9) > 0$, ….

[1]SSF : Splice Site Finder, Shapiro *et al., Nucleic Acids Res.* 1987 ; [2]MES : MaxEntScan, Yeo *et al., J. Comput. Biol.,* 2004

Calcul à partir d'une séquence complète

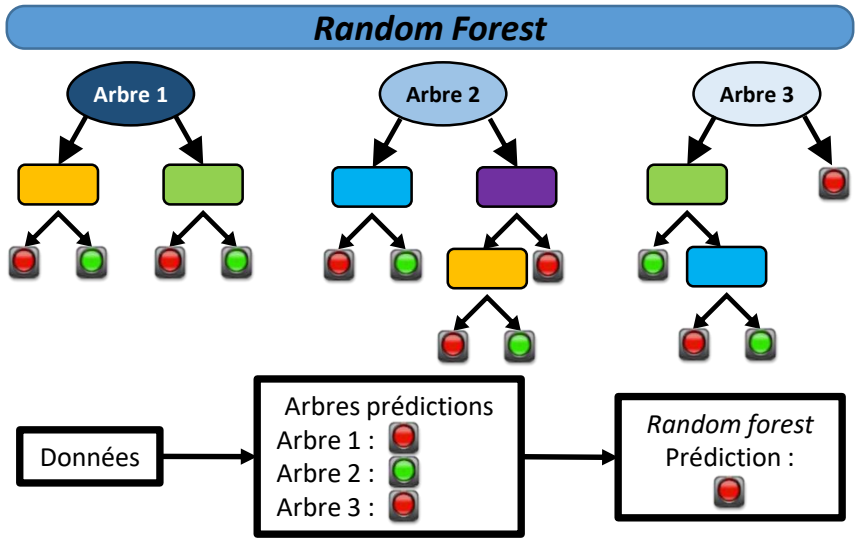
Grande diversité d'outils

Support Vector Machine (SVM)



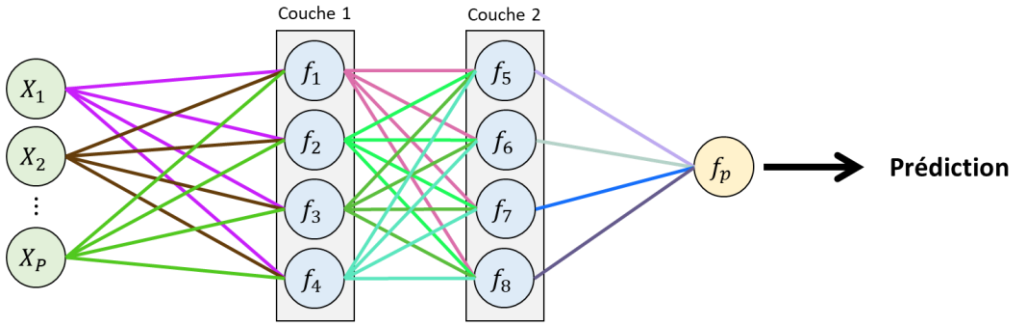
Ex : SVM-BPfinder; Corvelo, *PLOS Comput. Biol.*, 2010

Random Forest (RF)



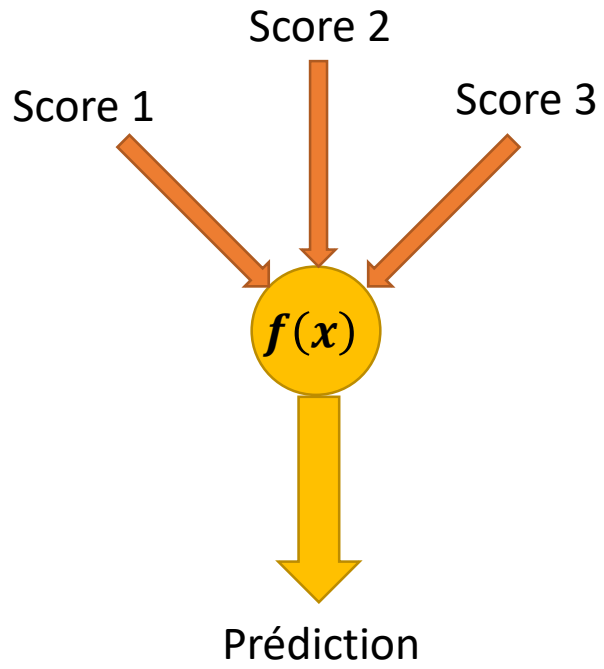
Ex : Meher *et al., BioData Min.*, 2016

Réseaux de neurones



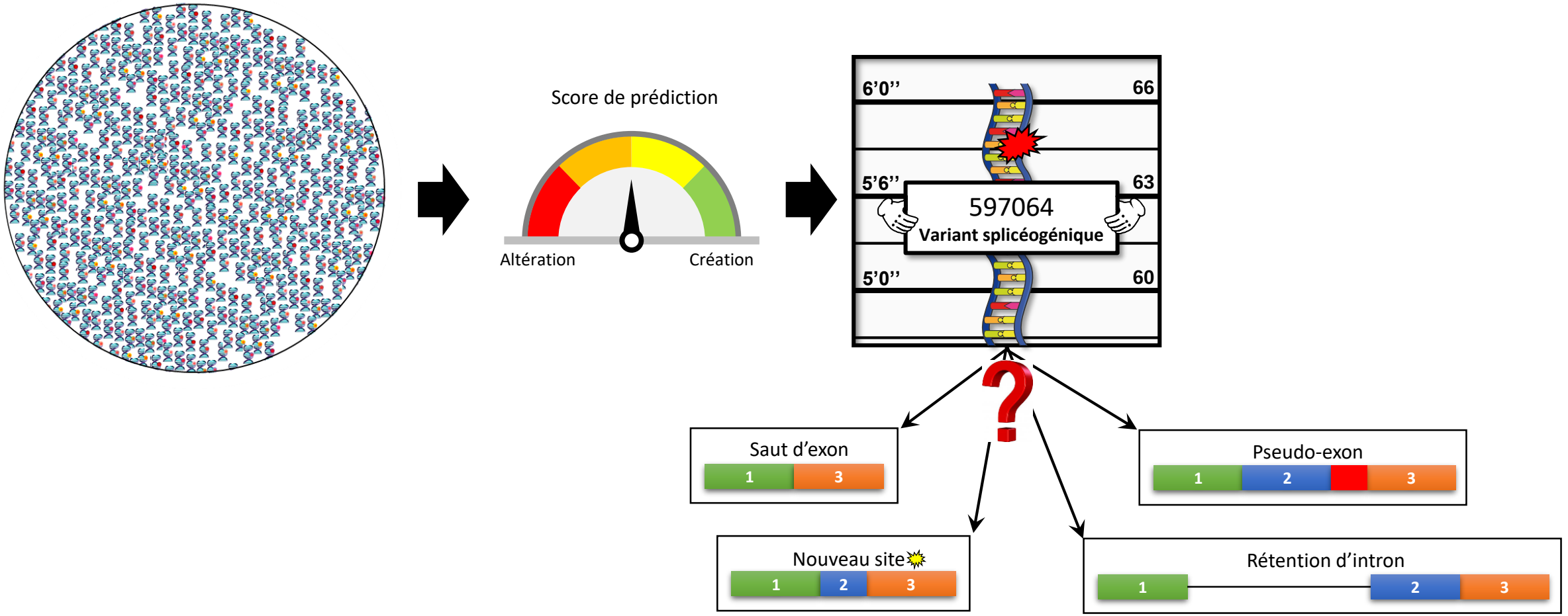
Ex : NNS (Neural Network Splice); Reese *et al.*, *In Gene-Finding and Gene Structure*, 1995

Meta-Score



Ex : ADAboost; Jian *et al., Nucleic Acids Res.*, 2014

Étude des variants splicéogéniques

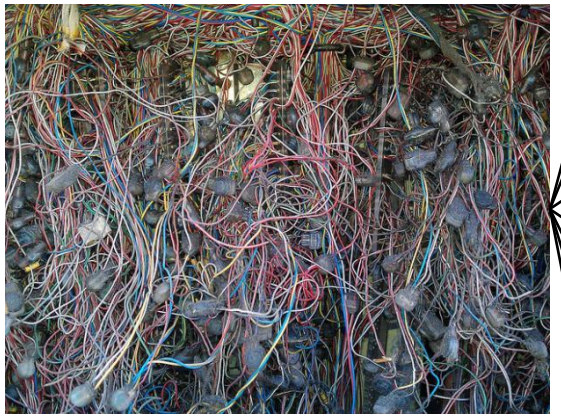


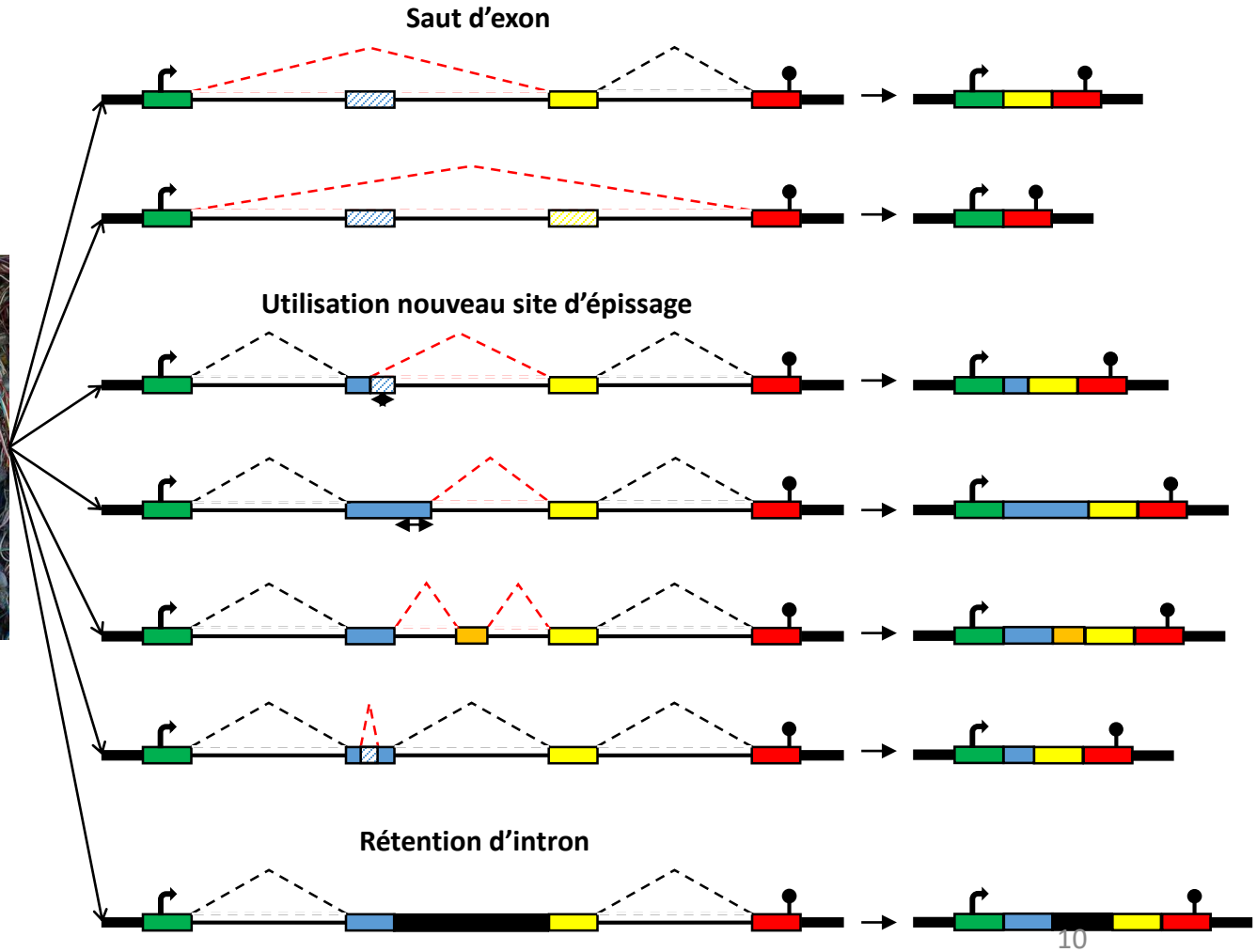
Épissage alternatif : rôle majeur dans la diversité des transcrits

À partir du même pré-mRNA

Épissage Naturel

Épissage alternatif





Humain

Gène : *KCNMA1*
> 500 ARNm ≠

Drosophile

Gène : *Dscam*
> 35 000 ARNm ≠

Schmucker *et al.*, *Cell*, 2000; Rosenblatt *et al.*, *Neuron*, 1997

Tests fonctionnels

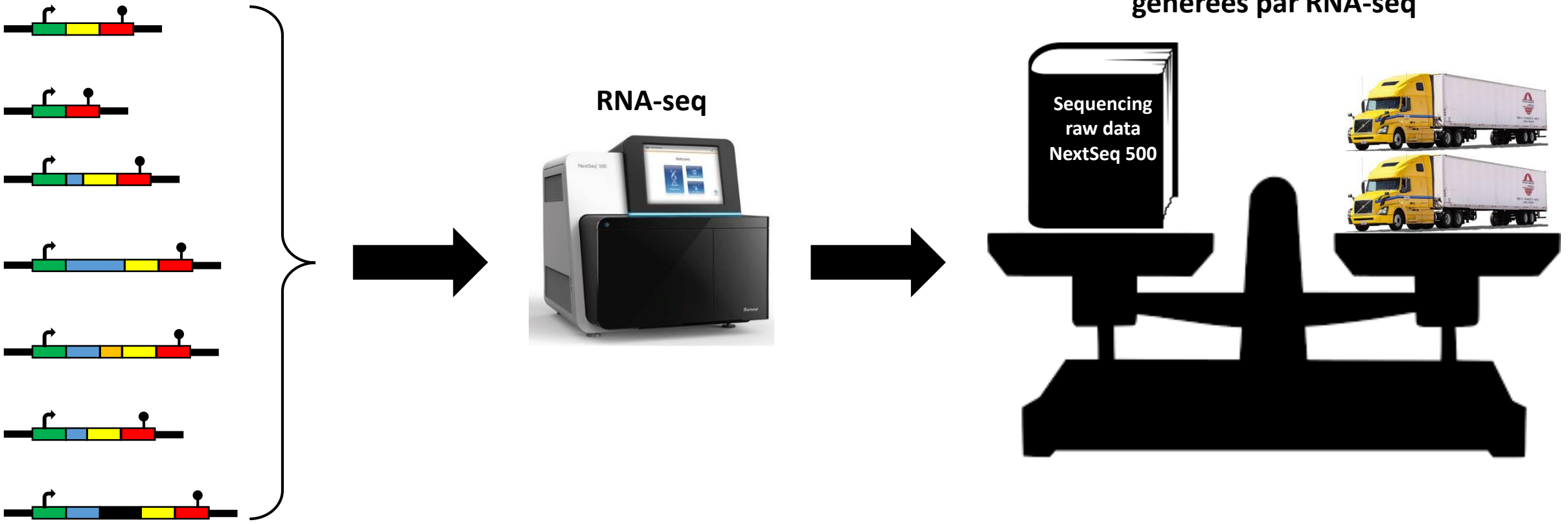
Tests à haut débit
RNA-seq

1 Gb – 1 Tb

Tests à bas débit
RT-PCR, test minigène, electrophorese capillaire, …

< 10 kb

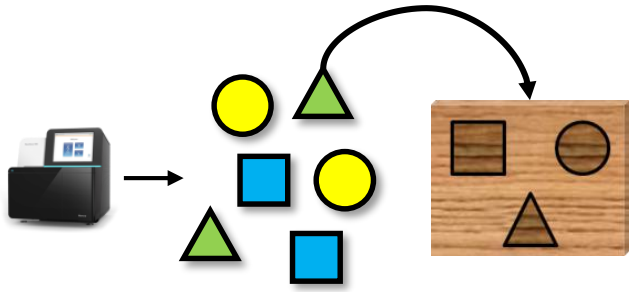
RNA-seq peut répondre à la diversité des transcrits
➔ Comment extraire la structure de ces transcrits

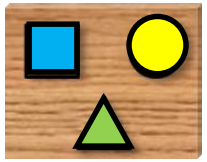


RNA-seq et analyses bioinformatiques

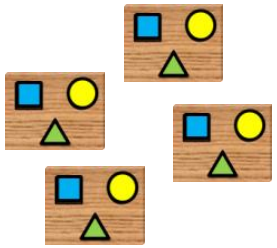
Alignement → Annotation → Quantification

- Visualisation des données
- Analyses biostatistiques



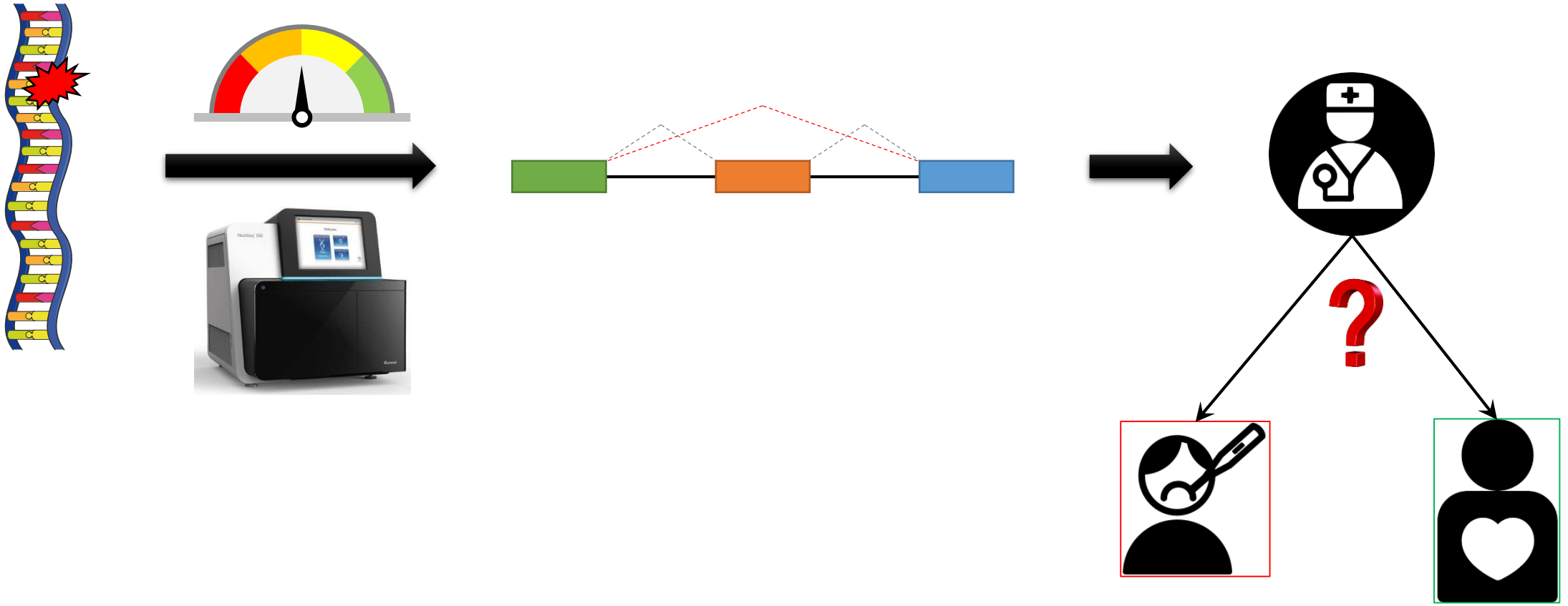


C’est un triangle vert
C’est un carré bleu
C’est un cercle jaune



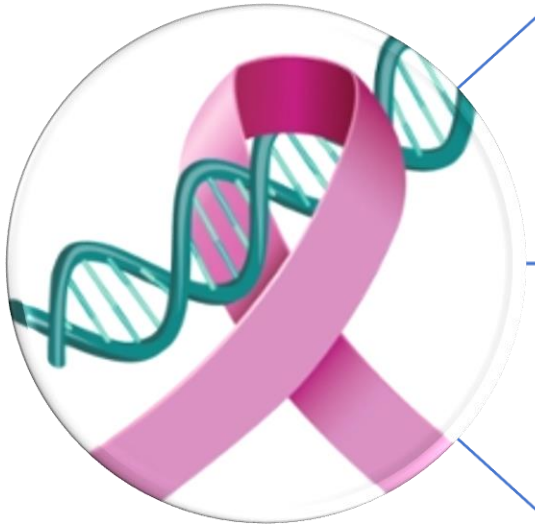
Combien de triangle par échantillon
Combien de carré par échantillon
Combien de cercle par échantillon

Transcrits et phénotypes



Épissage et pathogénicité

Étude de la prédisposition aux cancers du sein et de l'ovaire



Cancers fréquents

- Cancer du sein ⇔ Cancer le plus fréquent chez la femme
- Cancer de l'ovaire⇔ Cancer gynécologique le plus fréquent

Hérédité

- 5 à 10 % des cancers du sein et de l'ovaire ont une cause héréditaire

Expérience

- Plus ¼ de siècle de recherche de causes héréditaires
- Grande collection de variants

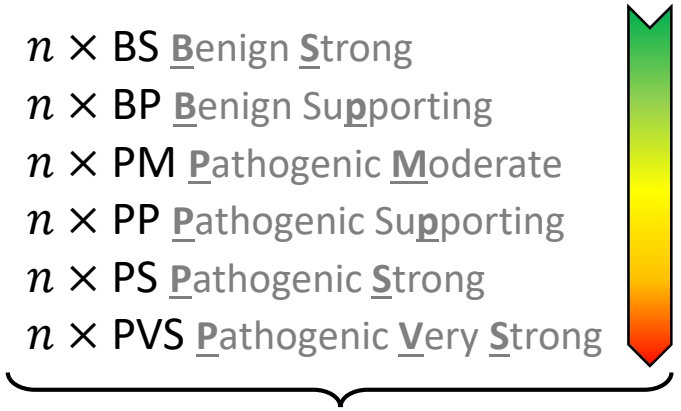
5 gènes majeurs de predisposition :

- *BRCA1* (1994)
- *BRCA2* (1995)
- *PALB2* (2007)
- *RAD51C* (2010)
- *RAD51D* (2011)

Classification des variants à usage clinique

ACMG
American College of Medical Genetics and Genomics

$n \times$ BS Benign Strong
$n \times$ BP Benign Supporting
$n \times$ PM Pathogenic Moderate
$n \times$ PP Pathogenic Supporting
$n \times$ PS Pathogenic Strong
$n \times$ PVS Pathogenic Very Strong



Algorithme décisionnel

Classe 1 | **Classe 2** | **Classe 3** | **Classe 4** | **Classe 5**

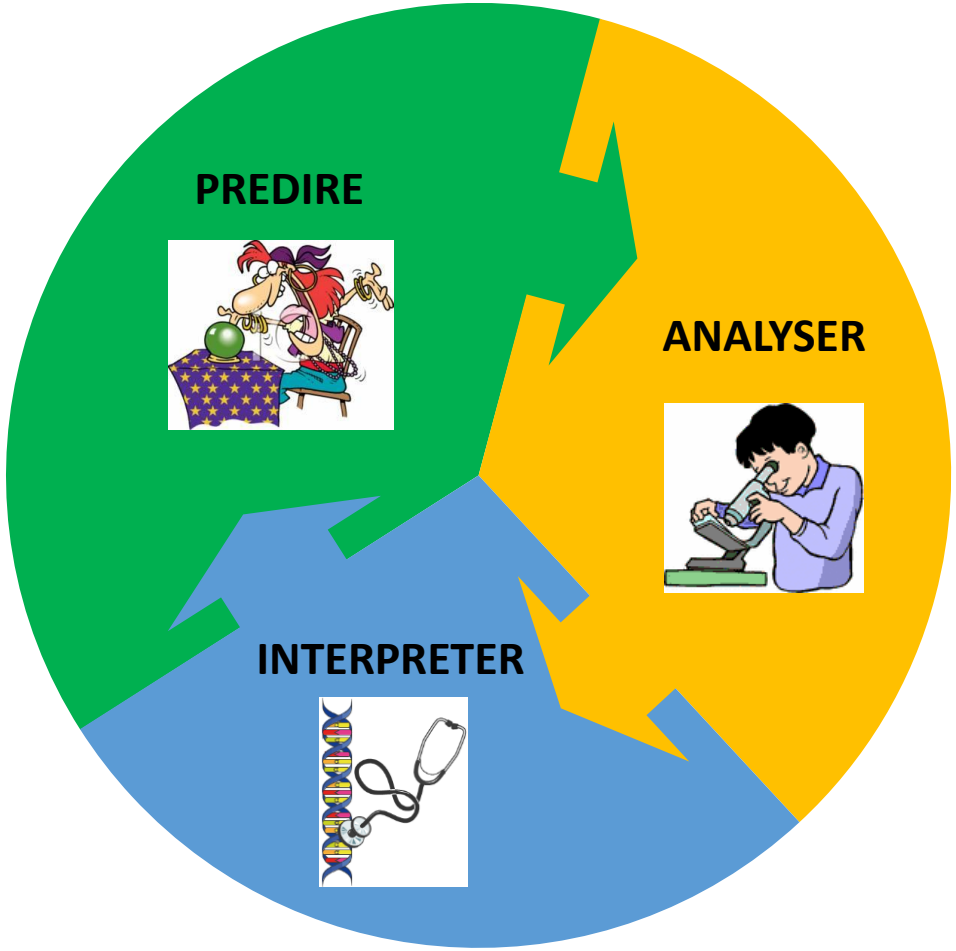
Classe 1	Neutre
Classe 2	Probablement neutre
Classe 3	VUS Variant of Unknown Significance
Classe 4	Probablement pathogène
Classe 5	Pathogène

Les **PVS** comprennent entre autres les variants **tronquants** et **canoniques** (-1/-2; +1/+2)

Richards, *et al., Genet. Med. Off. J. Am. Coll. Med. Genet.,* 2015

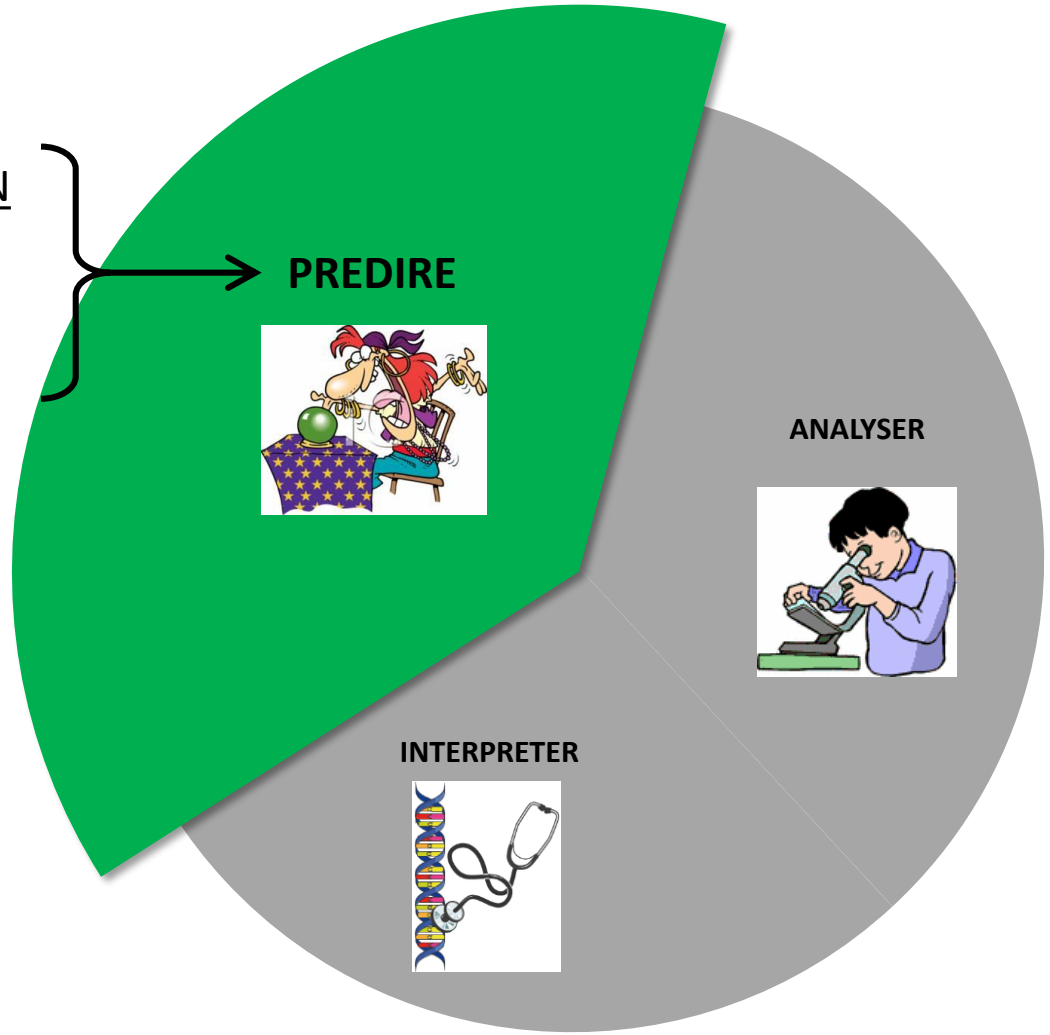
Objectif des travaux de thèse

Le cycle des études des défauts d'épissage

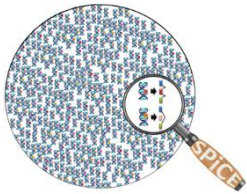


Sélection variants pour études ARN

- SPiCE
- Comparaison prédictions BP
- SPiP

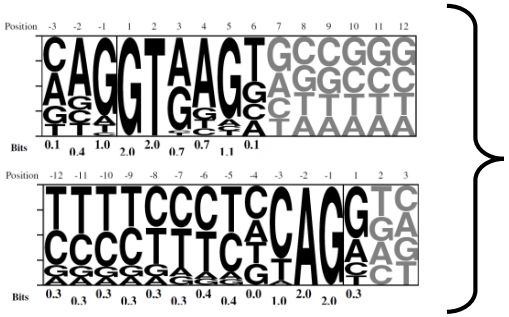


SPiCE, Splicing Prediction in Consensus Element *(1/6)*

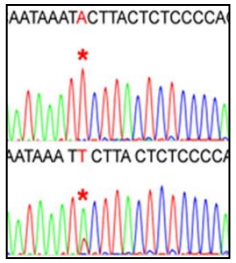


- **Objectif:**

  À partir des prédictions bioinformatiques, être capable de prioriser l'étude ARN des variants à haut risque d'altérer l'épissage

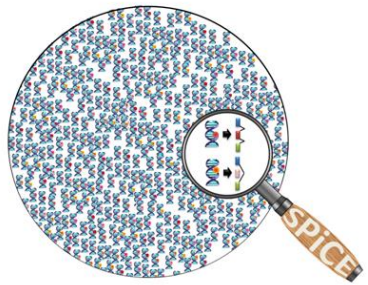


Mutation dans un site d'épissage consensus



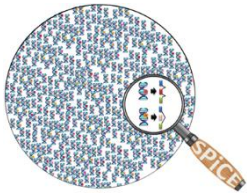
Scores de prédiction
MES, SSF-like*, HSF, NNS, GS

* Adaptation aux sites GC

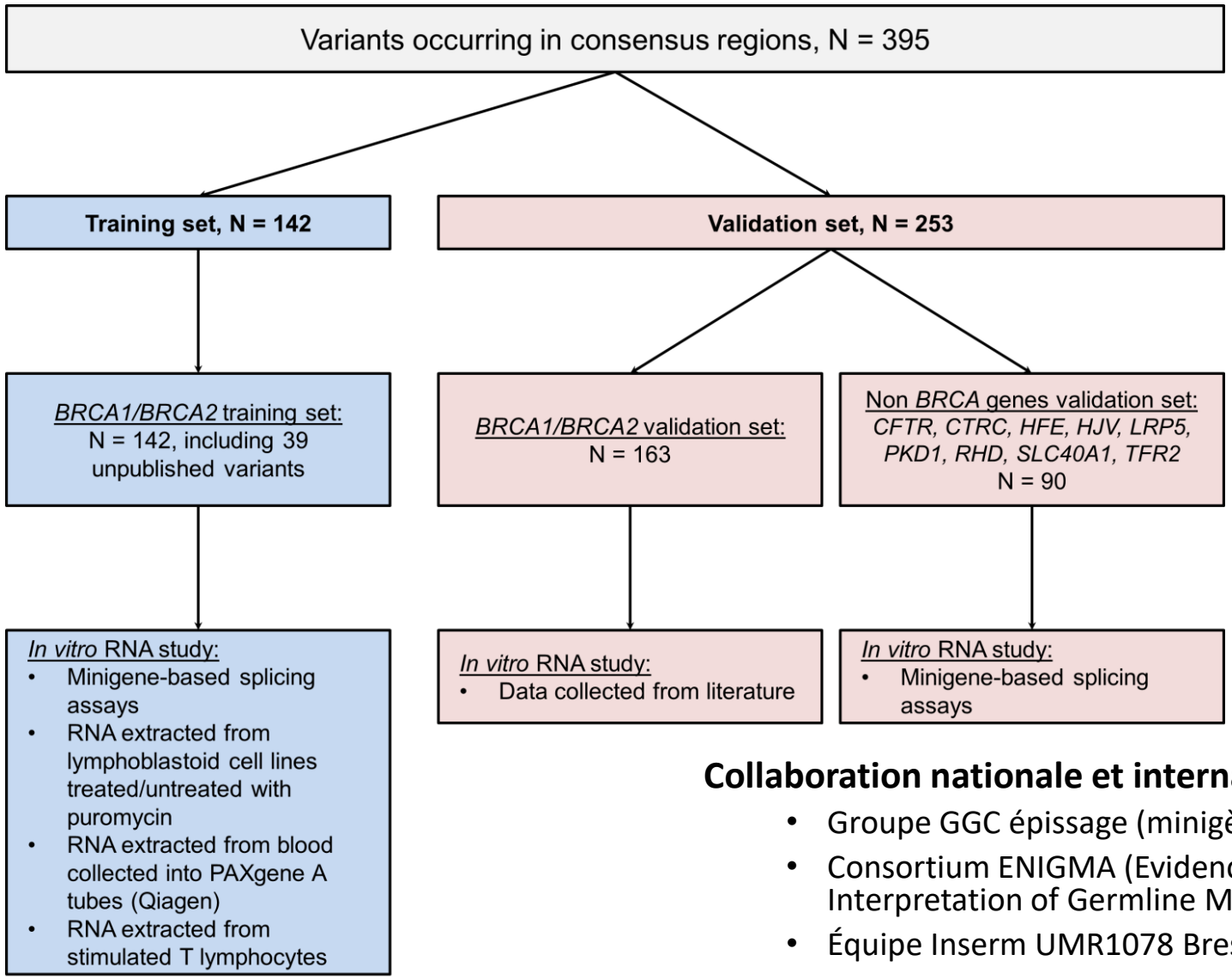


Pas d'étude ARN

Études ARN *in vitro*

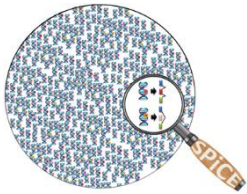


Jeux de données *(2/6)*



Collaboration nationale et internationale :

- Groupe GGC épissage (minigènes U1245, Rouen)
- Consortium ENIGMA (Evidence-based Network for the Interpretation of Germline Mutant Alleles)
- Équipe Inserm UMR1078 Brest



Construction de l'outil *(3/6)*

Données d'apprentissage
+
Régression logistique

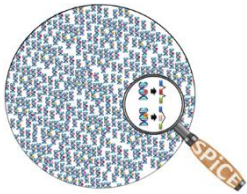
Variables utilisées

- Scores de prédictions (SSF-like, MES, HSF, NNS et GS)

Sélection variables

Validation modèle

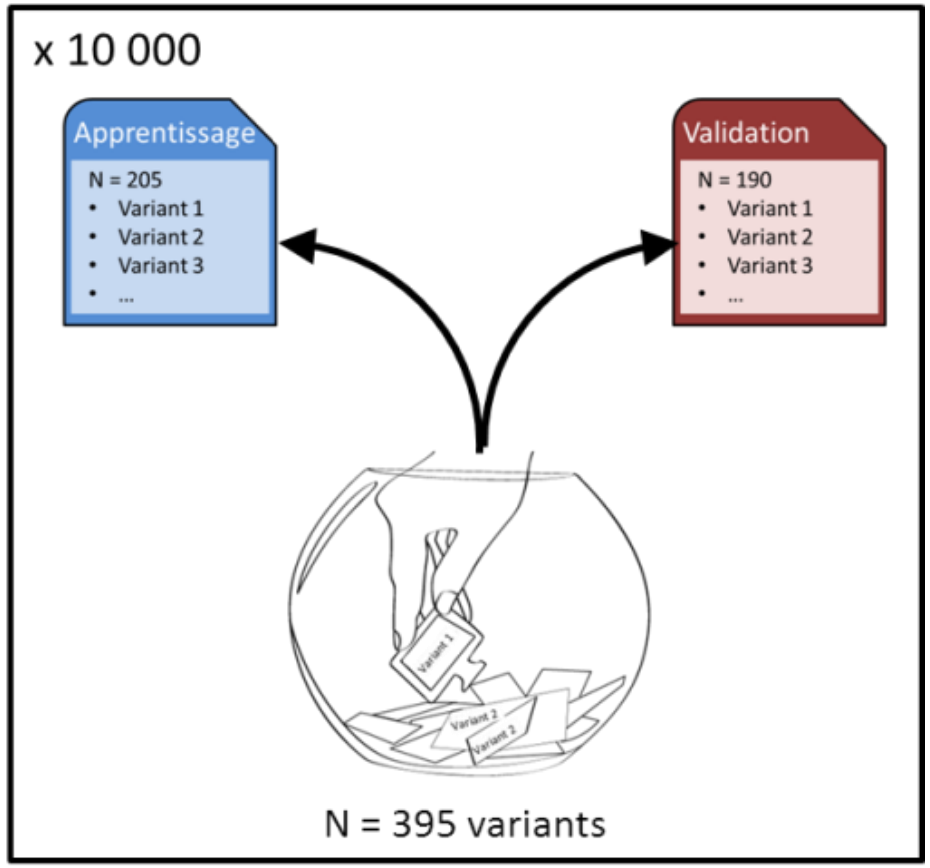
Modèle final + cross-validation

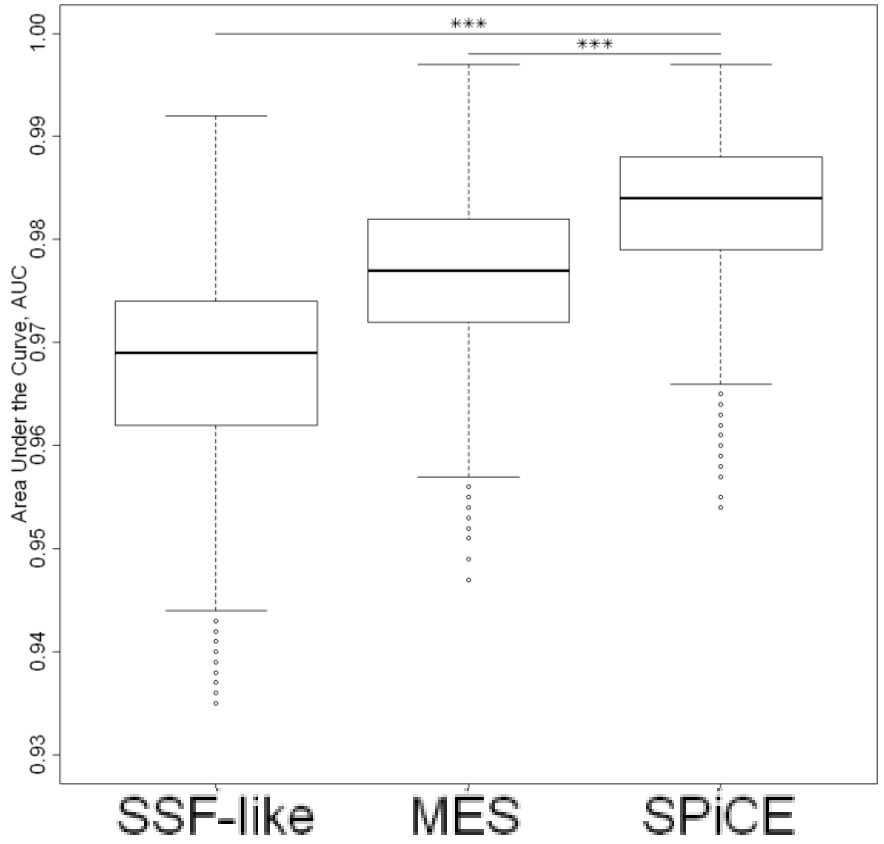


Validation de l'outil *(4/6)*

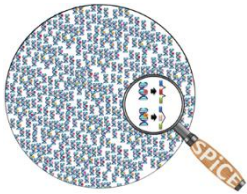
SPiCE : Combinaison par régression logistique des deux scores **MES** et **SSF-like**

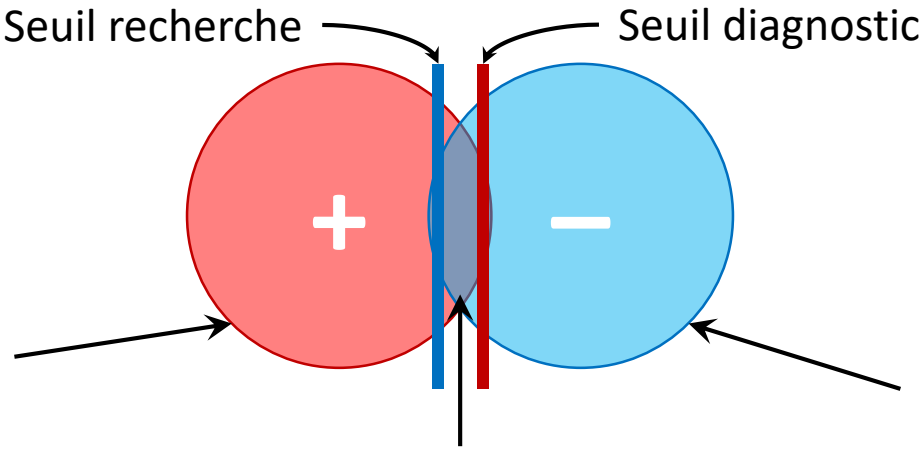
Cross-Validation





Performances et règles décisionnelles *(5/6)*





High

Score SPiCE **supérieur** au seuil =>
Études ARN *in vitro* recommandées

Low

Score SPiCE **inférieur** au seuil =>
Études ARN *in vitro* **ne sont pas** recommandées

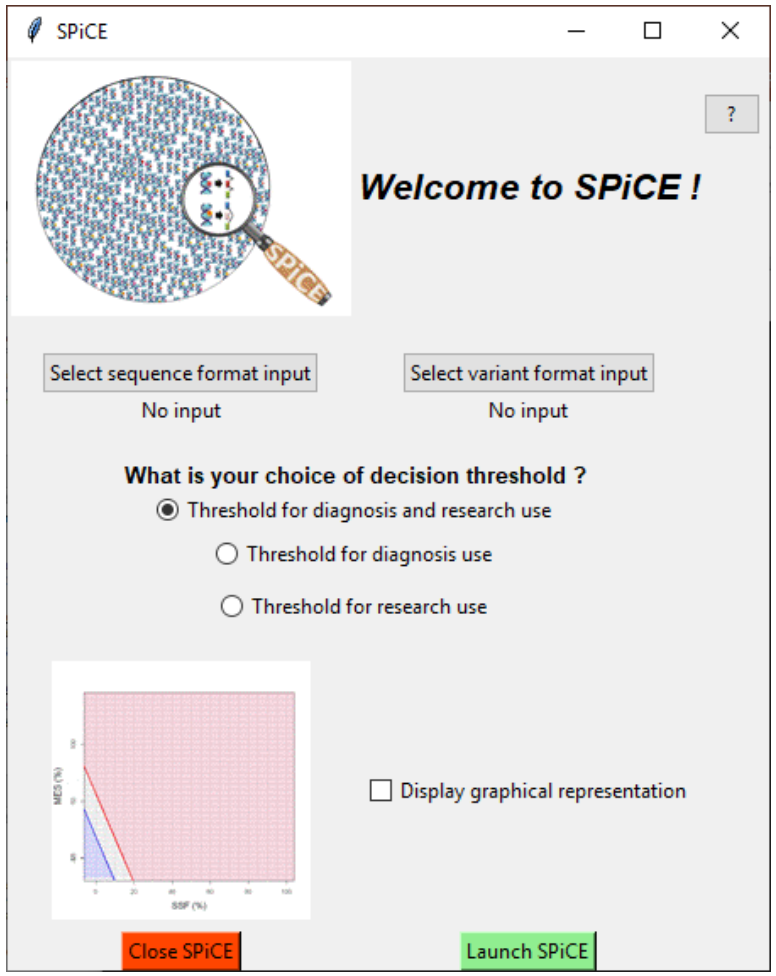
Medium (si les deux seuils sont utilisés)

'**zone grise**' => Etudes ARN *in vitro* en fonction du contexte clinique

Seuil adapté au diagnostic moléculaire :
Sensibilité = 99,5 %
Spécificité = 75,6 %

Seuil adapté à la recherche :
Sensibilité = 92,3 %
Spécificité = 95,6 %

Application et diffusion de SPiCE *(6/6)*



+

Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined *in silico/in vitro* studies: an international collaborative effort

Raphaël Leman, Pascaline Gaildrat, Gérald L Gac, Chandran Ka, Yann Fichou, Marie-Pierre Audrezet, Virginie Caux-Moncoutier, Sandrine M Caputo, Nadia Boutry-Kryza, Mélanie Léone, Sylvie Mazoyer, Françoise Bonnet-Dorion, Nicolas Sevenet, Marine Guillaud-Bataille, Etienne Rouleau, Brigitte Bressac-de Paillerets, Barbara Wappenschmidt, Maria Rossing, Danielle Muller, Violaine Bourdon, Françoise Revillon, Michael T Parsons, Antoine Rousselin, Grégoire Davy, Gaia Castelain, Laurent Castéra, Joanna Sokolowska, Florence Coulet, Capucine Delnatte, Claude Férec, Amanda B Spurdle, Alexandra Martins, Sophie Krieger ✉, Claude Houdayer ✉
Author Notes

Nucleic Acids Research, Volume 46, Issue 15, 6 September 2018, Pages 7913–7923, https://doi.org/10.1093/nar/gky372
Published: 10 May 2018 **Article history ▾**

Outils de prédiction des points de branchement (BP) *(1/5)*

Point de branchement
Court motif entre -44 et -18

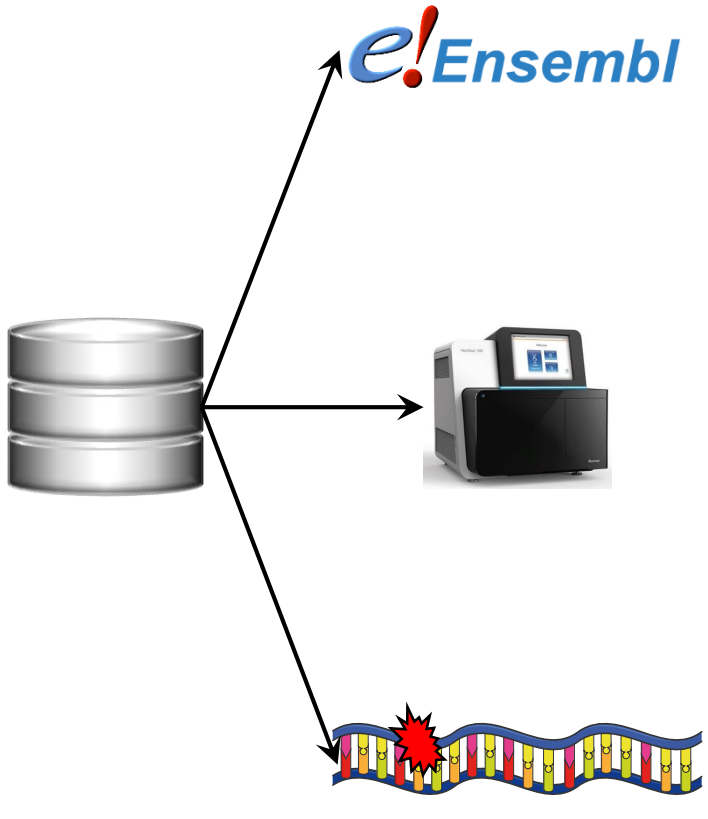
Pourquoi réaliser une comparaison des outils de prédiction des BPs ?

- **1977** — Découverte du mécanisme d'épissage[1]
- **1984** — 1er outil de prédiction, SSF[2]
- **2007** — 1er outil de prédiction BP, HSF[3]
- **2015** — 1er large collection de BPs expérimentalement prouvés[5]
- **2019** — SVM-BPfinder[4], BPP[6], Branchpointer[7], LaBranchoR[8], RNABPS[9]

6 outils sélectionnés : HSF, SVM-BPfinder, BPP, Branchpointer, LaBranchoR, RNABPS

[1]Chow *et al., Cell,* 1977
[2]Shapiro *et al., Nucleic Acids Res.,* 1987
[3]Desmet *et al., Nucleic Acids Res.,* 2009
[4]Corvelo *et al., PLOS Comput. Biol.,* 2010
[5]Mercer *et al., Genome Res.,* 2015
[6]Zhang *et al., Bionformatics*, 2017
[7]Signal *et al., Bioinformatics*, 2018
[8]Paggi *et al., RNA,* 2018
[9]Nazari *et al., IEEE Access,* 2019

Jeux de données et performances *(2/5)*

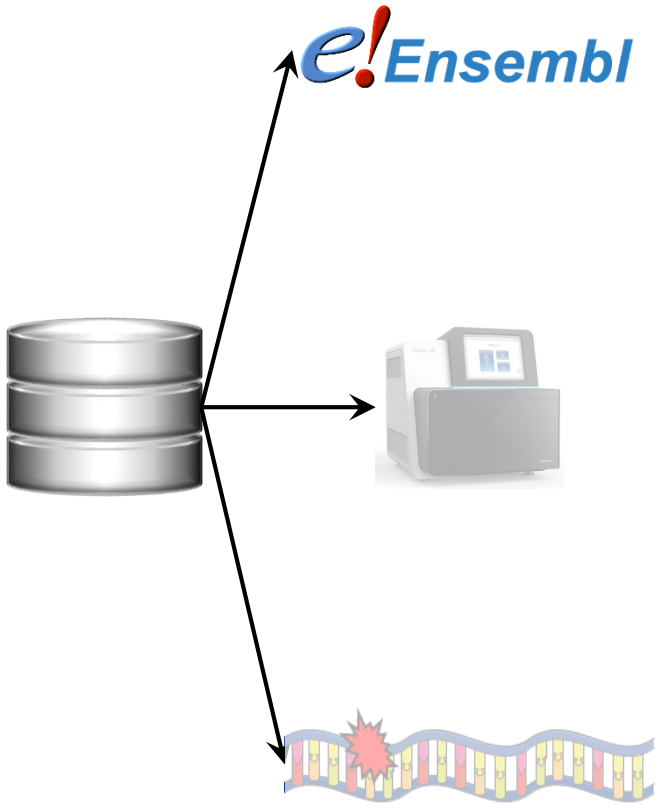


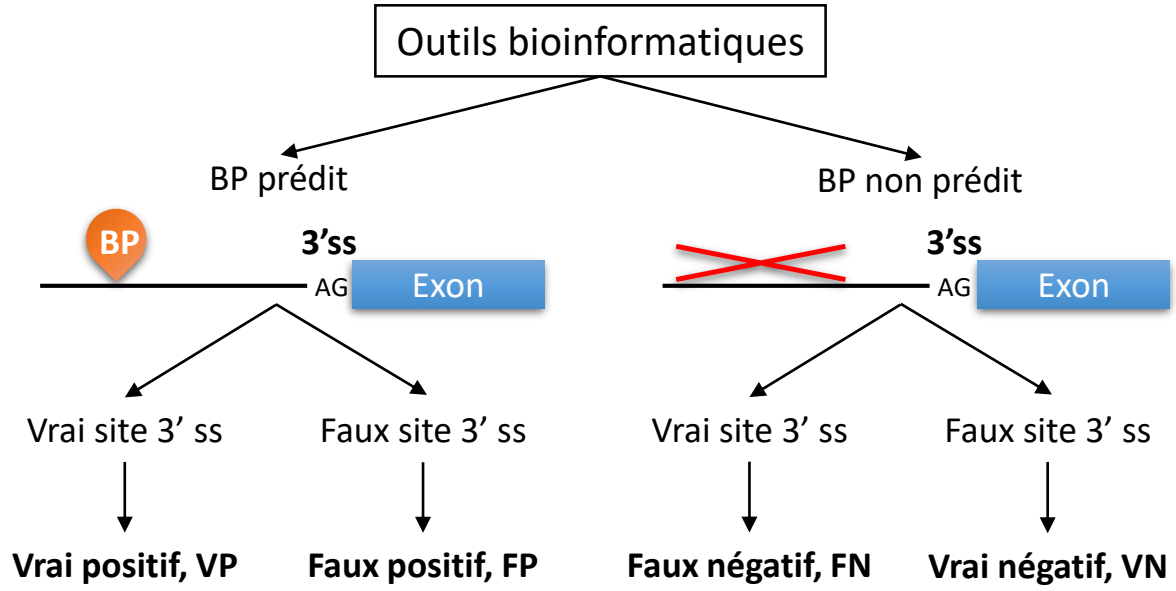
Sites physiologiques accepteurs
Sites négatifs tous motifs AG des 23 000 gènes

Sites accepteurs alternatifs par RNA-seq ➔ expression *versus* prédiction

Collection de variants avec étude ARN *in vitro* dans la région BP (-44 à -18)

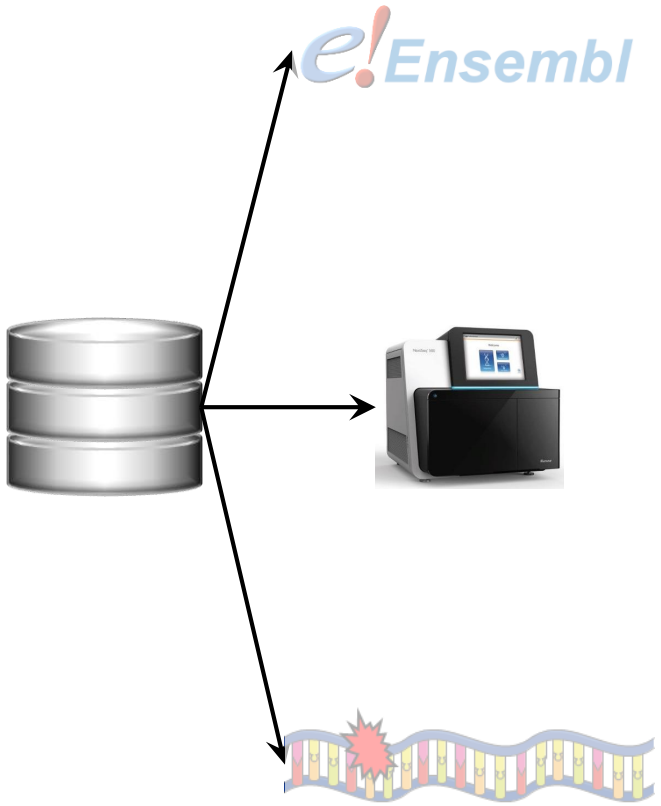
Jeux de données et performances *(3/5)*





	SVM-BPfinder	BPP	Branchpointer	LaBranchor	RNABPS
Exactitude	66.39 %	75.17 %	***99.48 %***	64.75 %	73.06 %
Sensibilité	66.39 %	75.17 %	***93.27 %***	64.75 %	73.05 %
Spécificité	66.39 %	75.17 %	***99.49 %***	64.75 %	73.06 %
VPP	0.45 %	0.70 %	***30.06 %***	0.43 %	0.62 %
VPN	99.88 %	99.92 %	***99.98 %***	99.87 %	99.91 %

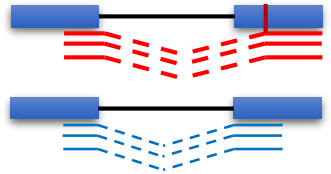
Jeux de données et performances *(3/5)*

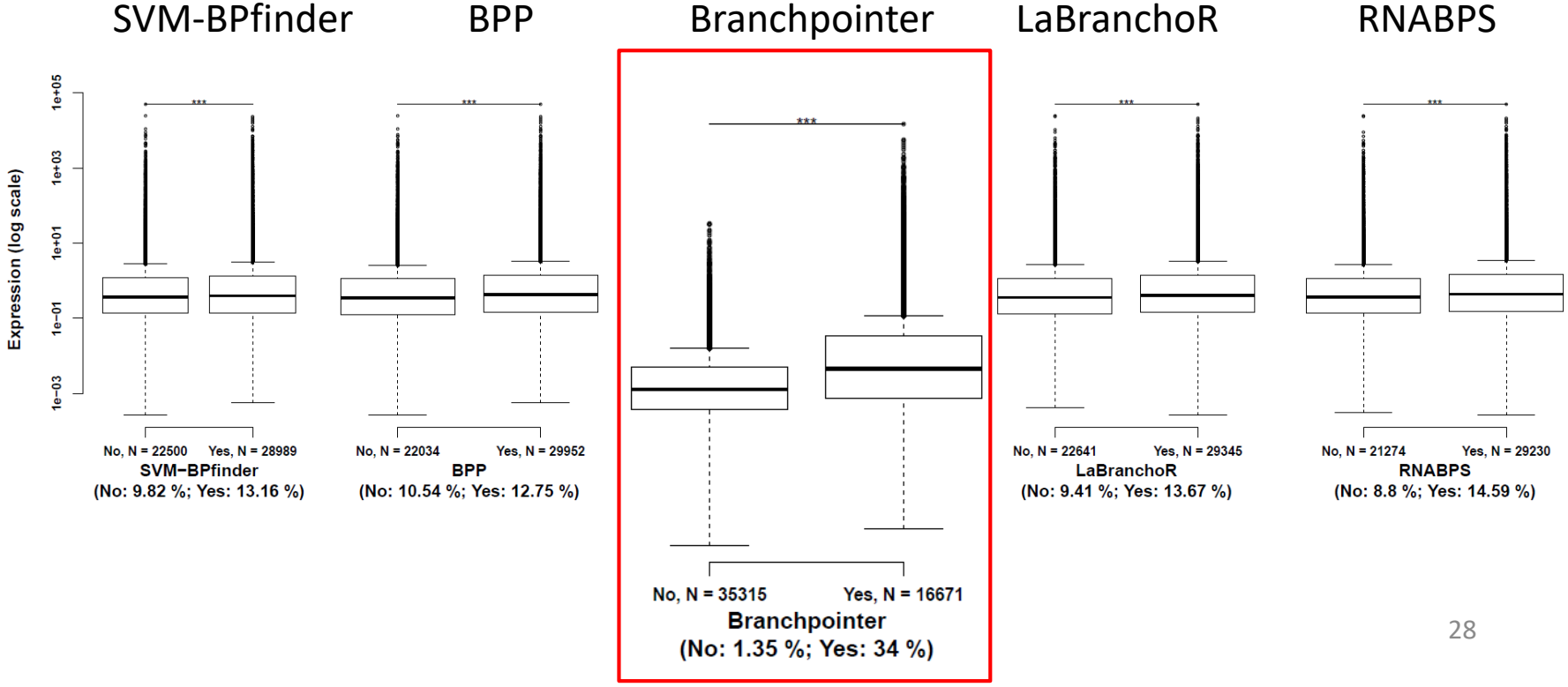


N = 72 échantillons ARN (LCLs)

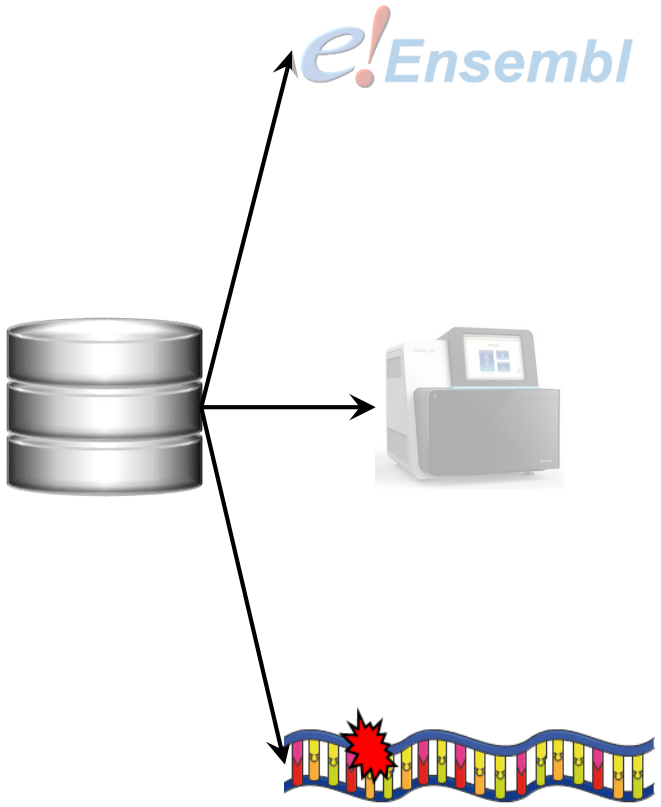
NexSeq 500

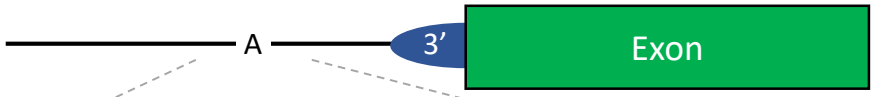
$$\%_{expression} = \frac{nb\ read\ 3'alternatif}{nb\ read\ 3'physiologique}$$





Jeux de données et performances *(3/5)*



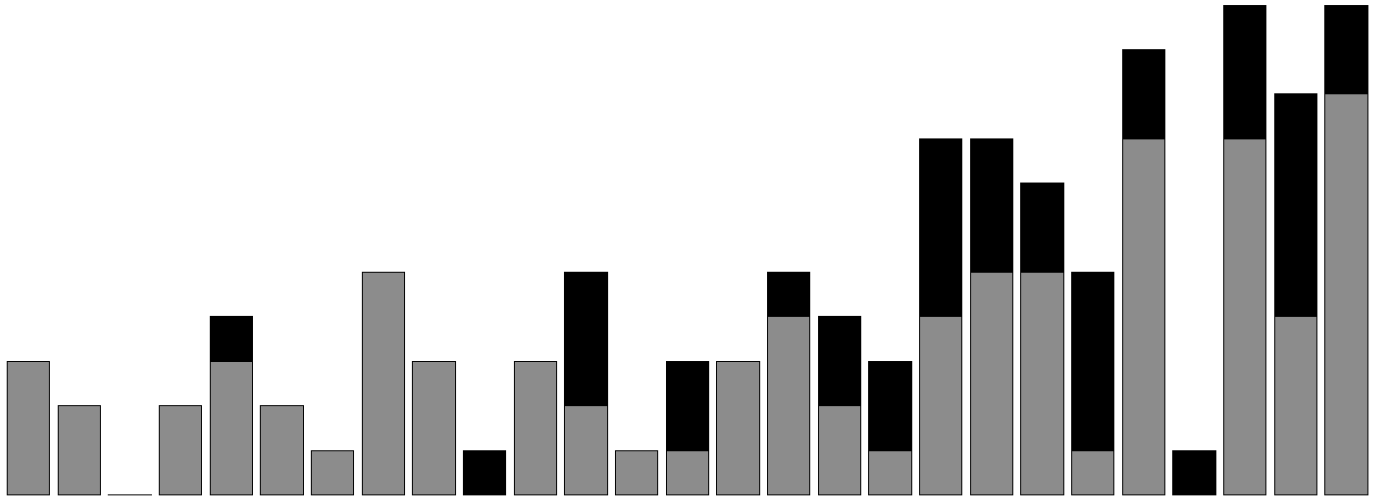


N = 120 variants

Région point de branchement

Nombre de variants pour chaque position:

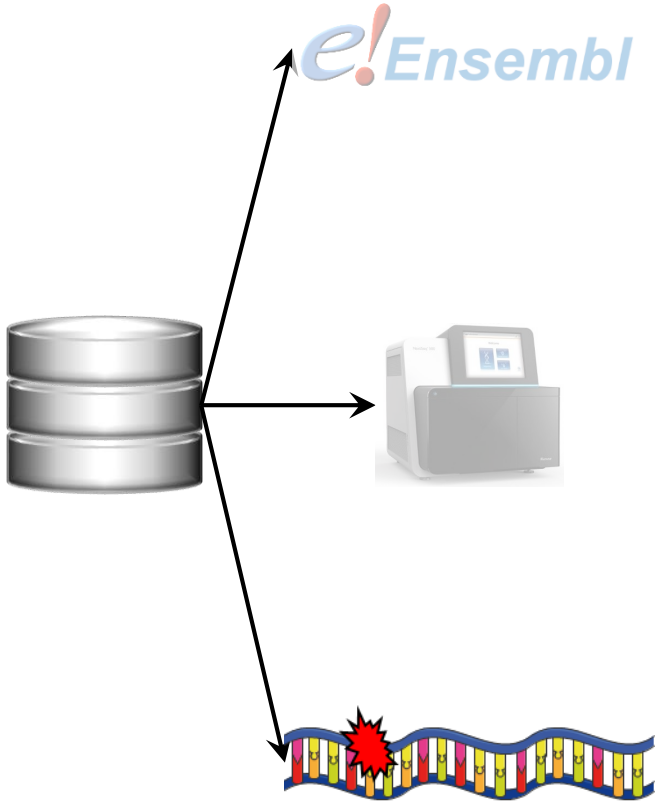
−44	−43	−42	−41	−40	−39	−38	−37	−36	−35	−34	−33	−32	−31	−30	−29	−28	−27	−26	−25	−24	−23	−22	−21	−20	−19	−18
0	0	0	0	1	0	0	0	0	1	0	3	0	2	0	1	2	2	4	3	2	4	2	1	3	5	2
3	2	0	2	3	2	1	5	3	0	3	2	1	1	3	4	2	1	4	5	5	1	8	0	8	4	9



Modifie l’épissage

Sans effet

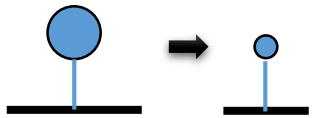
Jeux de données et performances *(3/5)*



- Est-ce que le variant est situé dans un point de branchement ?

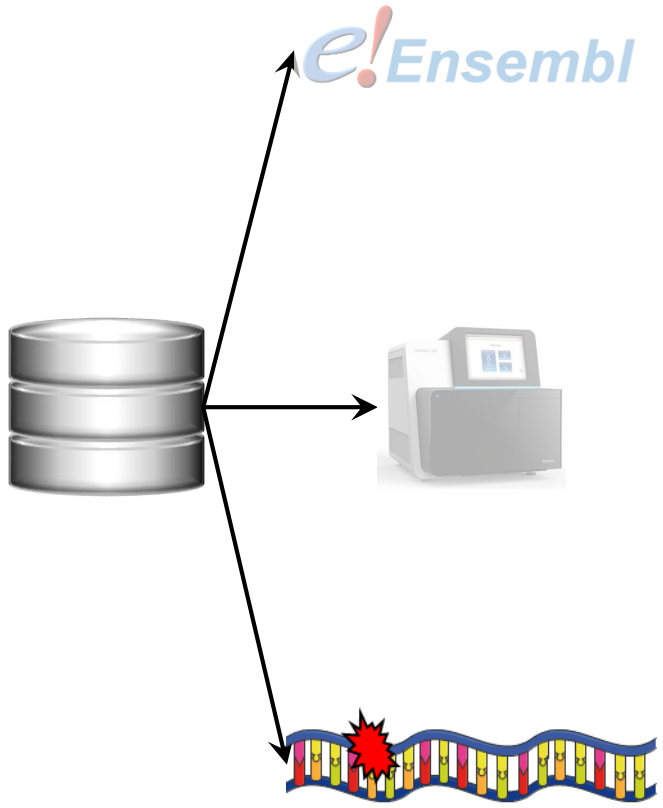
TRAY *

- Est-ce que le variant diminue le score du point de branchement ?



*taille du motif (4-mers) déterminée pour avoir le meilleur compromis sensibilité/spécificité

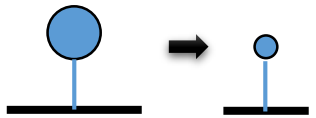
Jeux de données et performances *(3/5)*



- Est-ce que le variant est situé dans un point de branchement ?

TRAY *

- Est-ce que le variant diminue le score du point de branchement ?



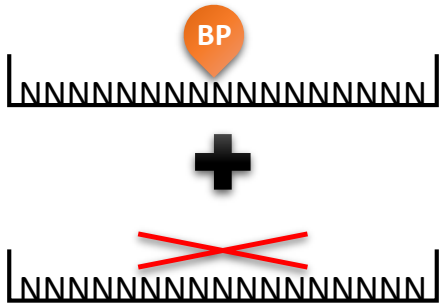
	SVM-BPfinder	BPP	Branchpointer	LaBranchoR	RNABPS
FP	6	7	12	15	12
FN	14	6	6	11	8
Exactitude	83.33 %	***89.17 %***	84.87 %	78.33 %	83.33 %
Sensibilité	63.16 %	***84.21 %***	84.21 %	71.05 %	78.95 %
Spécificité	***92.68 %***	91.46 %	85.19 %	81.71 %	85.37 %

*taille du motif (4-mers) déterminée pour avoir le meilleur compromis sensibilité/spécificité

Importance de l'apprentissage *(4/5)*

Détection d'un point de branchement parmi le bruit de fond

Branchpointer



VS

Autres outils

Altération d'un motif du point de branchement

BPP

Apprentissage sur les motifs les plus conservés sur plus de **200000 introns**

VS

Autres outils

Apprentissage sur des BPs putatifs ou expérimentaux

Conclusion *(5/5)*

I. Pouvoir utiliser les prédictions des points de branchement pour expliquer l'utilisation ou non d'un site accepteur

- Meilleur score : **Branchpointer**

II. Prédire l'effet d'un variant intronique sur un point de branchement

- Meilleur score : **BPP**

RESEARCH ARTICLE *Epigenetics & Genomics*

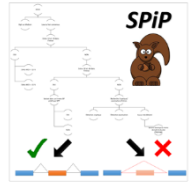
Assessment of branch point prediction tools to predict physiological branch points and their alteration by variants

Raphael Leman, Hélène Tubeuf, Sabine Raad, Isabelle Tournier, Céline Derambure, Raphaël Lanos, Pascaline Gaildrat, Gaia Castelain, Julie Hauchard, Audrey Killian, Stéphanie Baert-Desurmont, Angelina Legros, Nicolas Goardon, Céline Quesnelle, Agathe Ricou, Laurent Castera, Dominique Vaur, Gérald Le Gac, Chandran Ka, Yann Fichou, Françoise Bonnet-Dorion, Nicolas Sevenet, Marine Guillaud-Bataille, Nadia Boutry-Kryza, Ines Schultz, Virginie Caux-Moncoutier, Maria Rossing, Logan C Walker, Amanda B Spurdle, Claude Houdayer, Alexandra Martins, Sophie Krieger

STATUS: IN REVISION

BMC Genomics

SPiP, Splicing Prediction Pipeline *(1/4)*



- **Objectif:**

  Développer un outil de prédiction combinant les scores optimaux pour chaque motif

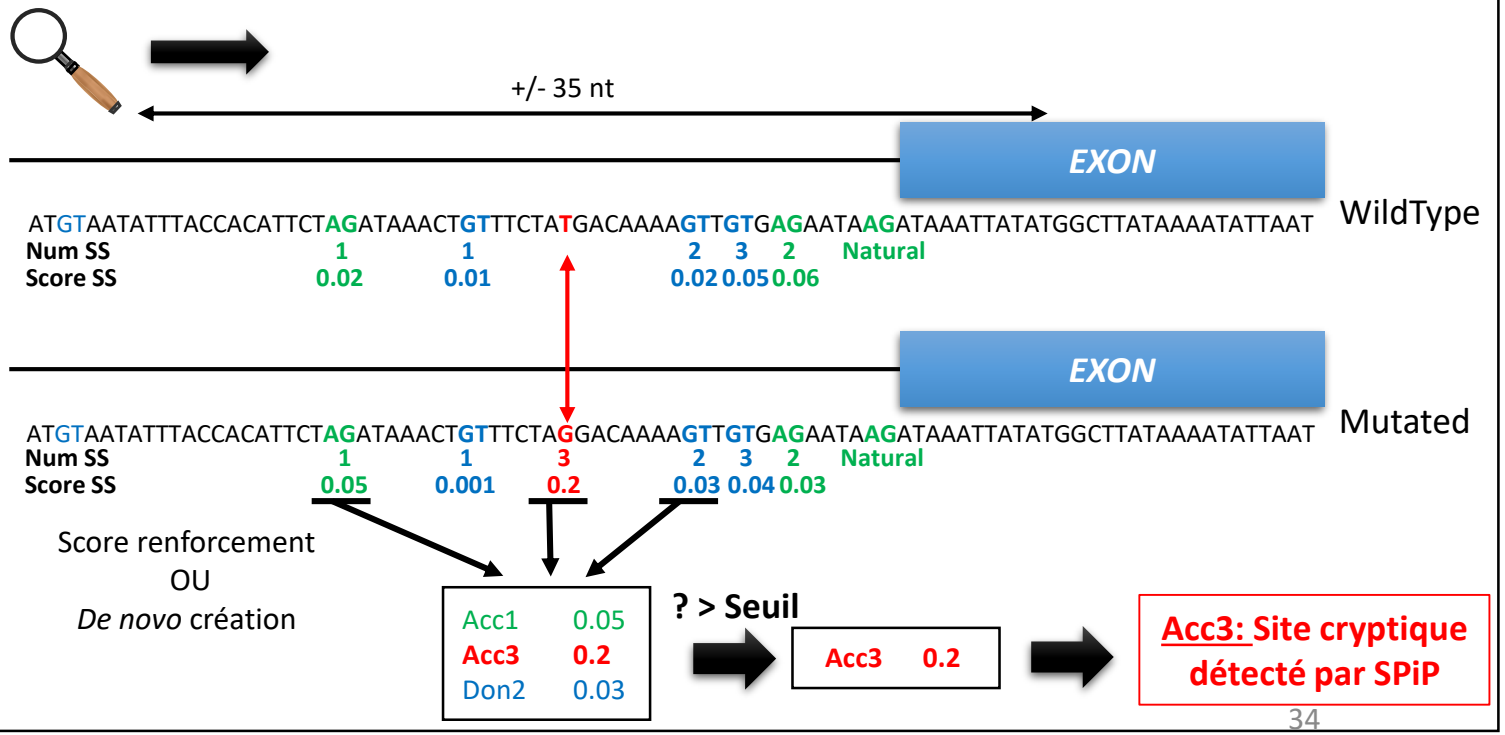
Ensembl + Control AG/GT

Comparaison des outils dans la littérature

- ✓ **SPiCE** ⇔ consensus sites[1]
- ✓ **BPP** ⇔ branch points[2]
- ✓ **ΔtESRseq** ⇔ESRs[3]
- ✓ **MES** ⇔ tract polypyrimidic[4]

+

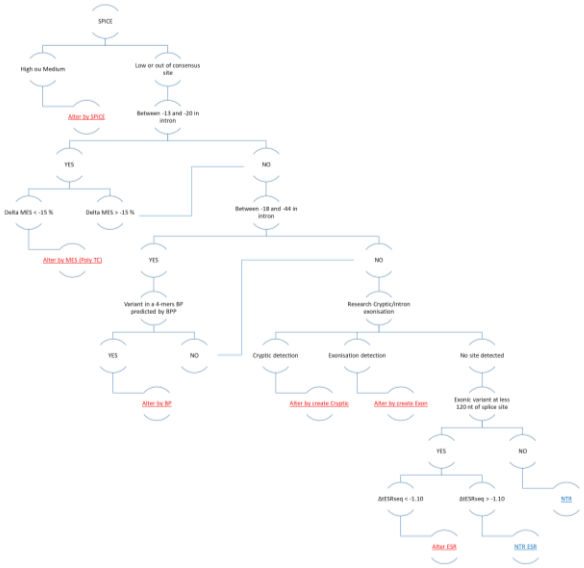
Nouvel outil pour la création d'un site d'épissage

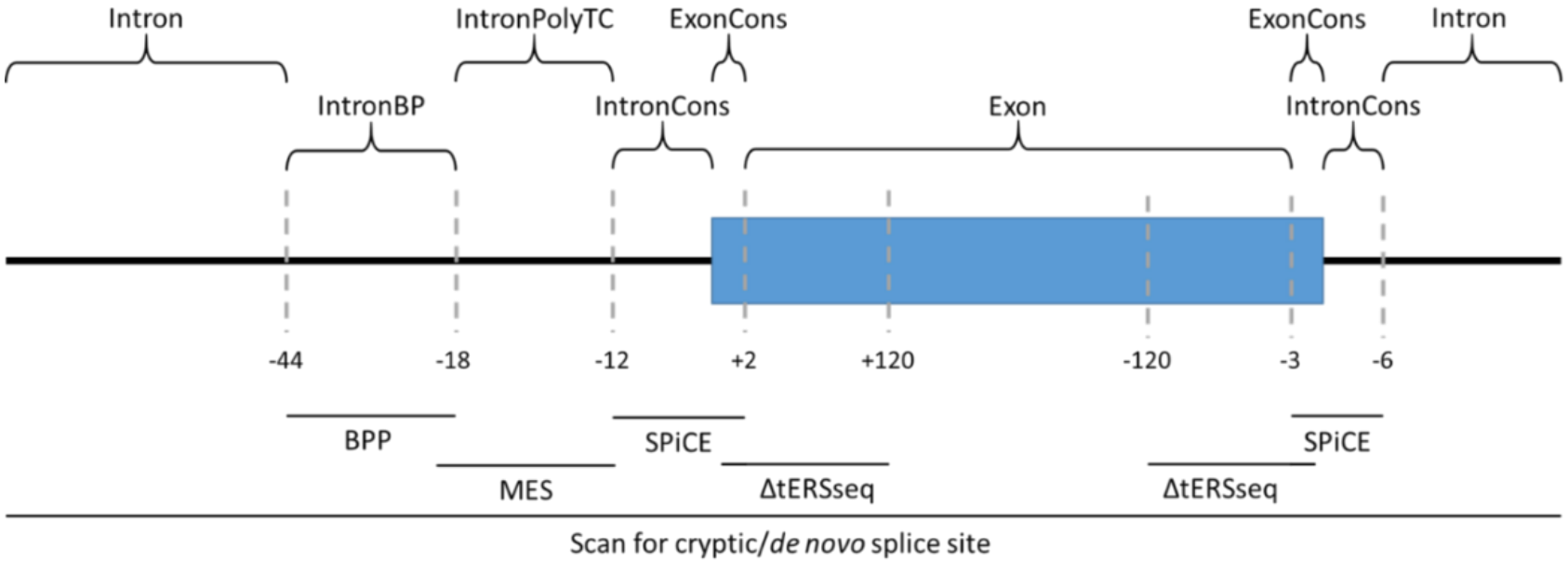


[1]Leman *et al., Nucleic Acids Res.*, 2018; [2]Leman *et al., BMC Genomics*, In revision; [3]Soukarieh *et al., PLOS Genet.*, 2016; [4]Houdayer *et al., Hum. Mutat.*, 2012

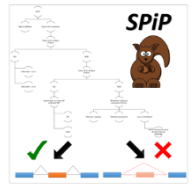
SPiP, Splicing Prediction Pipeline *(2/4)*

SPiP, **S**plicing **P**redi**c**tion **P**ipeline *(2/4)*

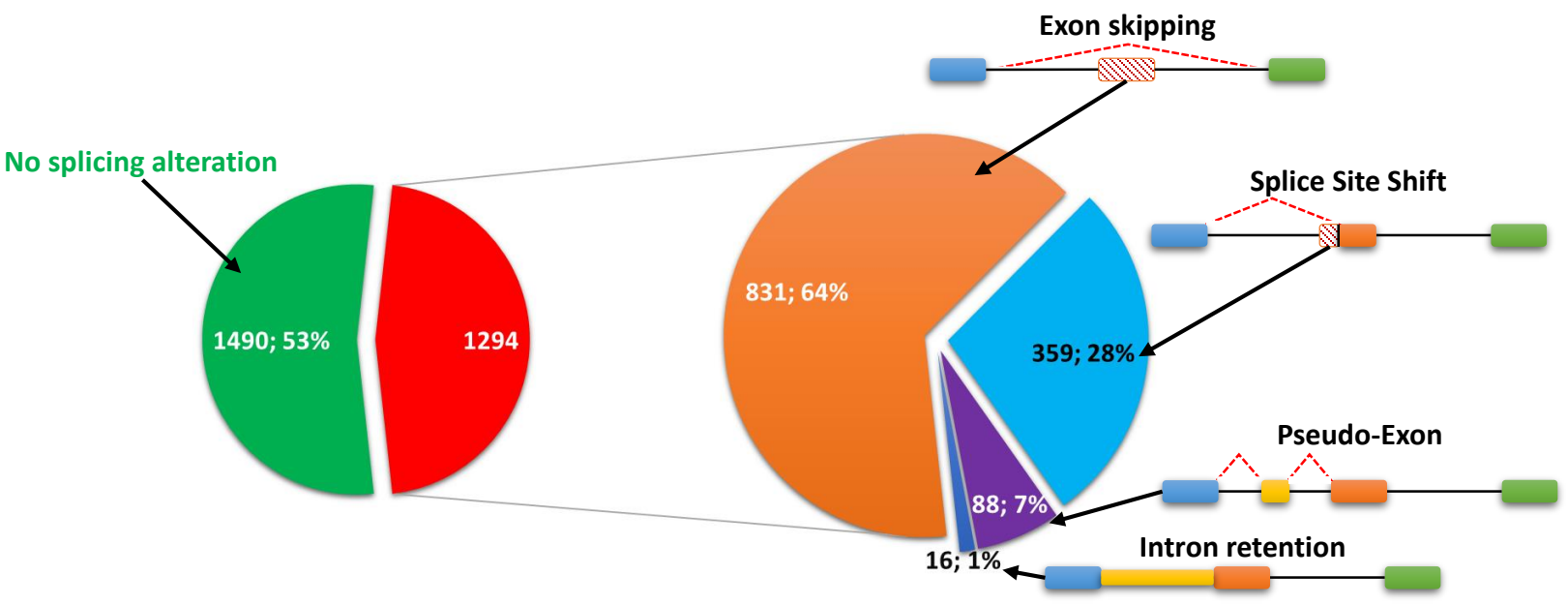




SPiP, Splicing Prediction Pipeline *(3/4)*



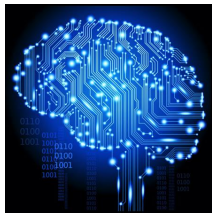
Jeux de 2 784 variants pour tester SPiP



SPiP

VS

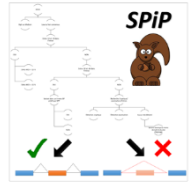
Deep learning



	EXACTITUDE	SENSIBILITÉ	SPÉCIFICITÉ
SPiP	**80.21 %**	**90.96 %**	**70.87 %**
SPANR[1]	75.45 %	78.37 %	72.32 %
SpliceAI[2]	85.45 %	70.71 %	98.26 %

[1]Xiong *et al., Science,* 2015 ; [2]Jaganathan *et al., Cell,* 2019

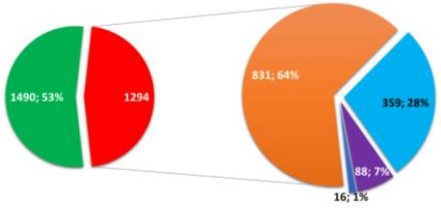
SPiP, Splicing Prediction Pipeline *(4/4)*



Estimation des probabilités d'altération de l'épissage selon :

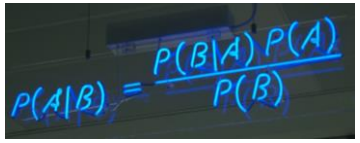
1) Position du variant
2) SPiP prédiction

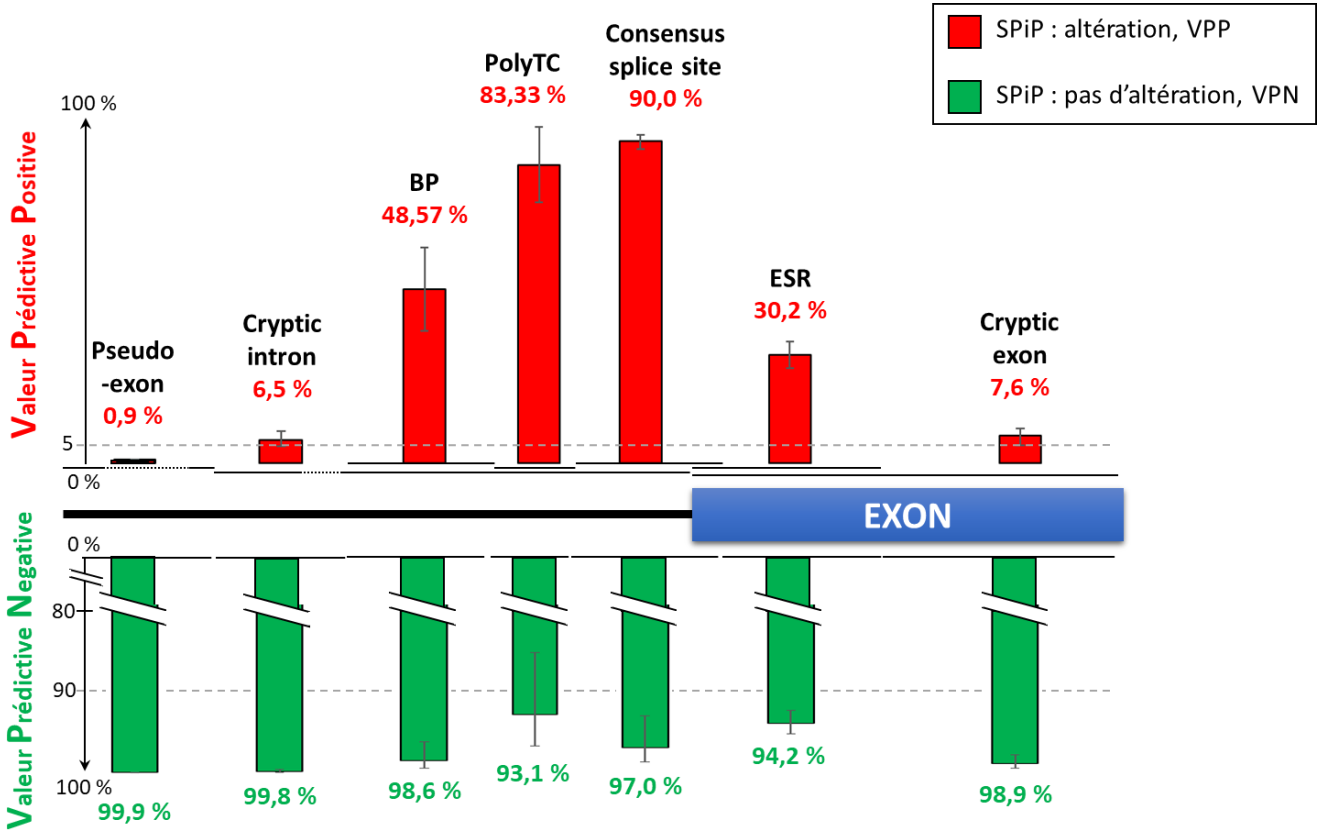
Le jeu de 2 784 variants



Polymorphisme comme données contrôles (MAF > 5 %)
N = 55 487 SNPs

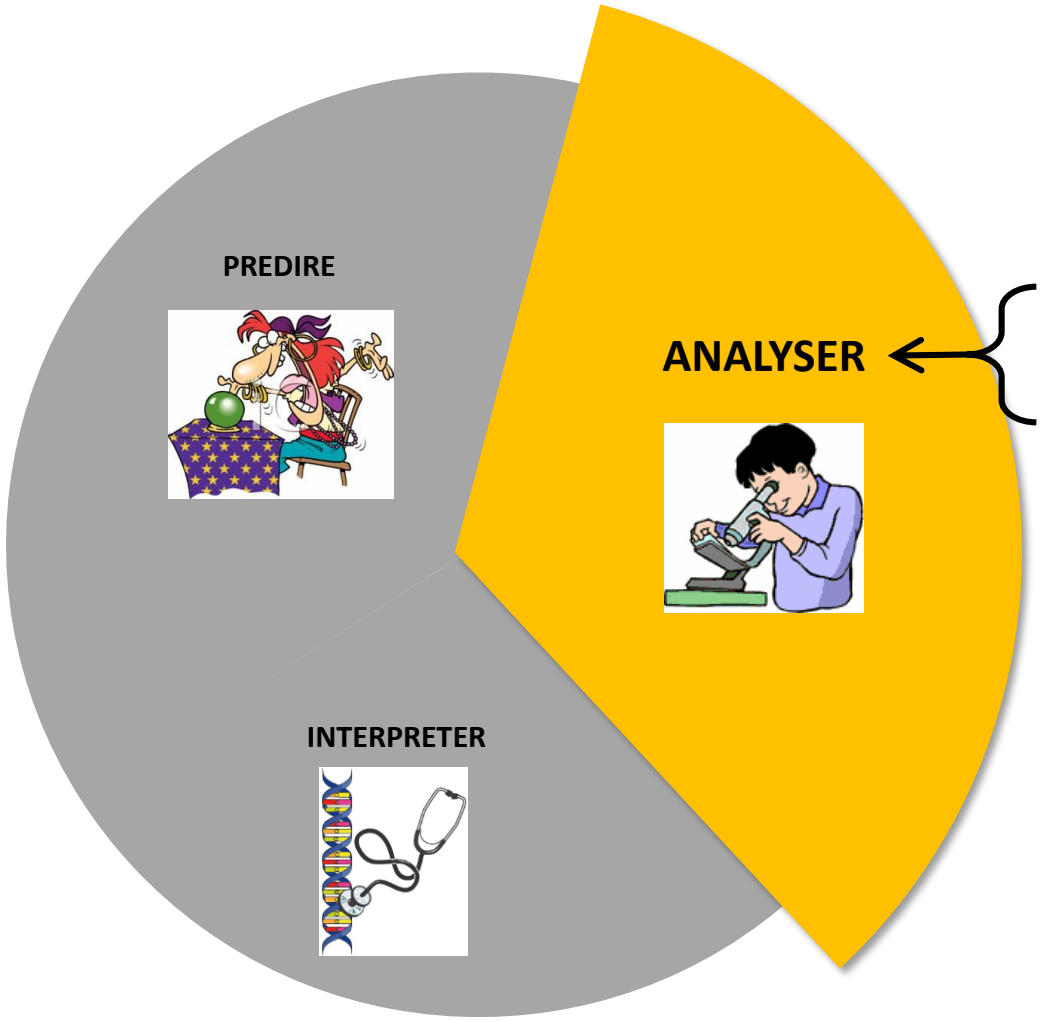
Théorème de Bayes





SPiP

Article en cours de rédaction, *SPiP: a Splicing Prediction Pipeline addressing the diversity of splice alterations, validated on a curated diagnostic set of 2,784 exonic and intronic*



Études *in vitro* : RNA-seq

- SpliceLauncher

SpliceLauncher *(1/4)*

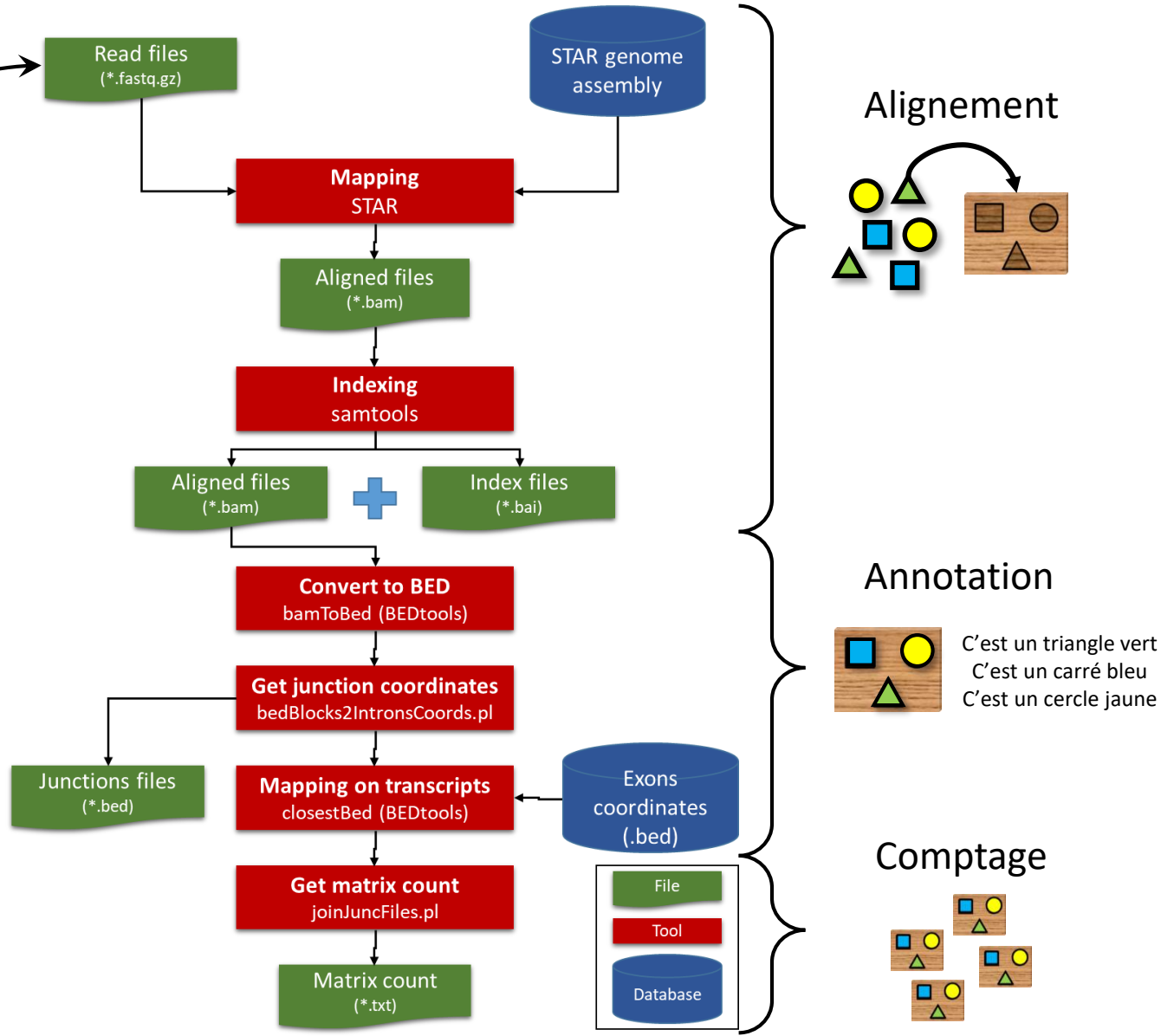
Plusieurs outils pour étudier l'expression des gènes et des exons **MAIS**
peu d'outils pour étudier la diversité des jonctions d'épissage

RNA-seq

SpliceLauncher

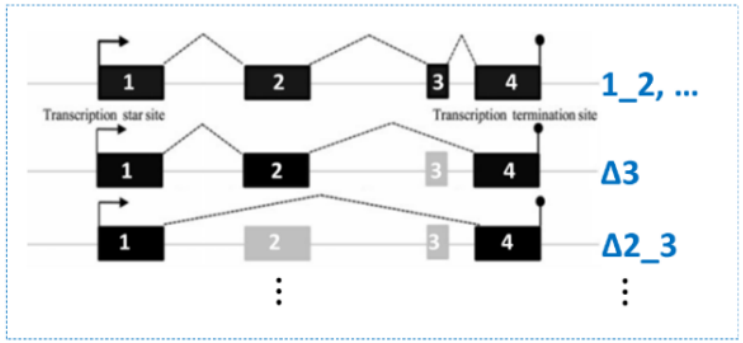
- Visualiser les jonctions d'épissage
- Annotation des épissages alternatifs
- Quantification relative
- Détection des jonctions anormalement exprimées

SpliceLauncher *(2/4)*

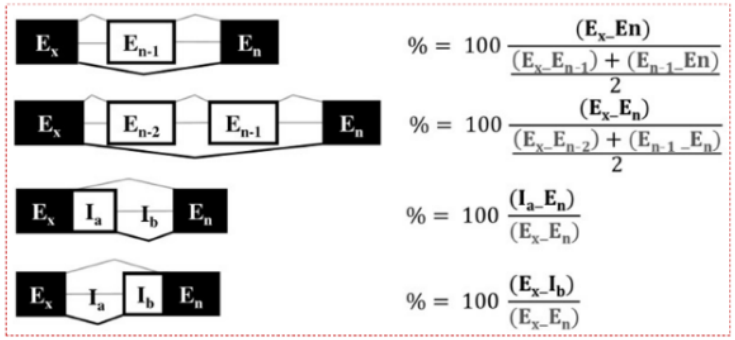


SpliceLauncher *(3/4)*

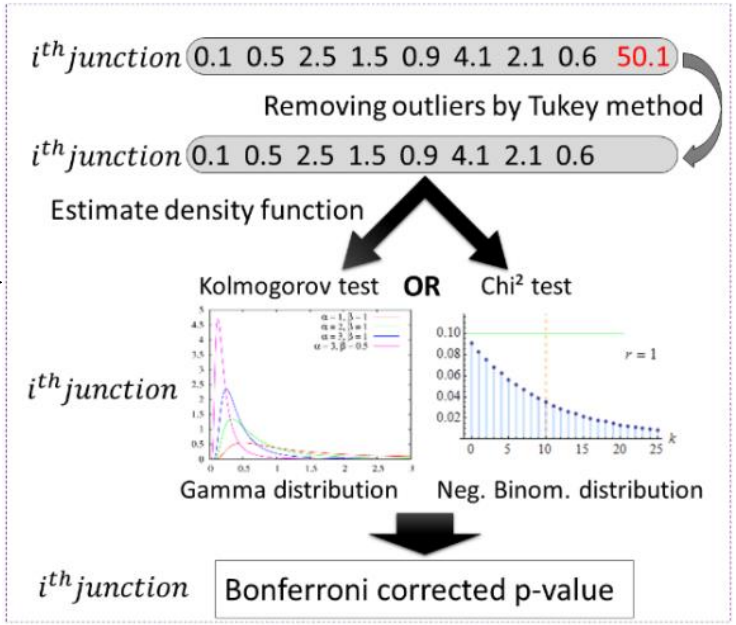
Annotation jonction



Expression relative



Analyse statistique



SpliceLauncher *(4/4)*

https://github.com/raphaelleman/SpliceLauncher

CORRECTED PROOF

SpliceLauncher: a tool for detection, annotation and relative quantification of alternative junctions from RNAseq data

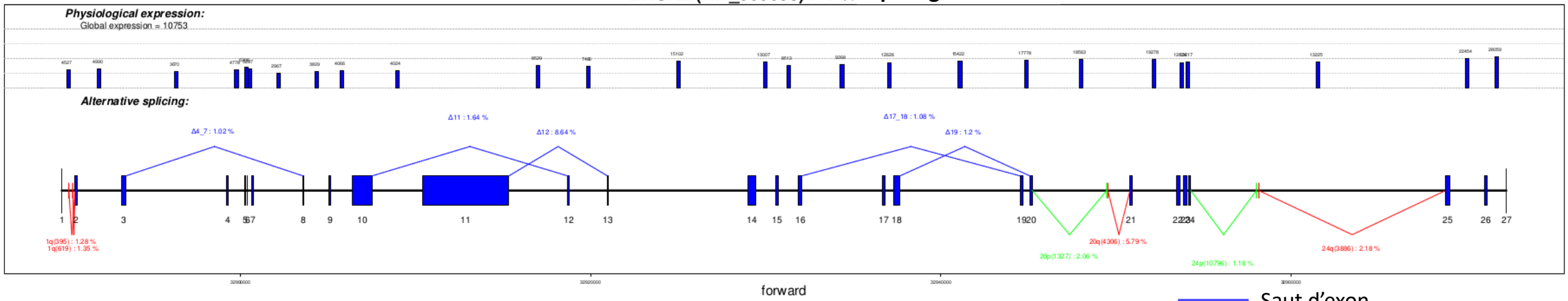
Raphaël Leman, Valentin Harter, Alexandre Atkinson, Grégoire Davy, Antoine Rousselin, Etienne Muller, Laurent Castéra, Fréderic Lemoine, Pierre de la Grange, Marine Guillaud-Bataille, Dominique Vaur, Sophie Krieger

Bioinformatics, btz784, https://doi.org/10.1093/bioinformatics/btz784
Published: 16 October 2019 **Article history**

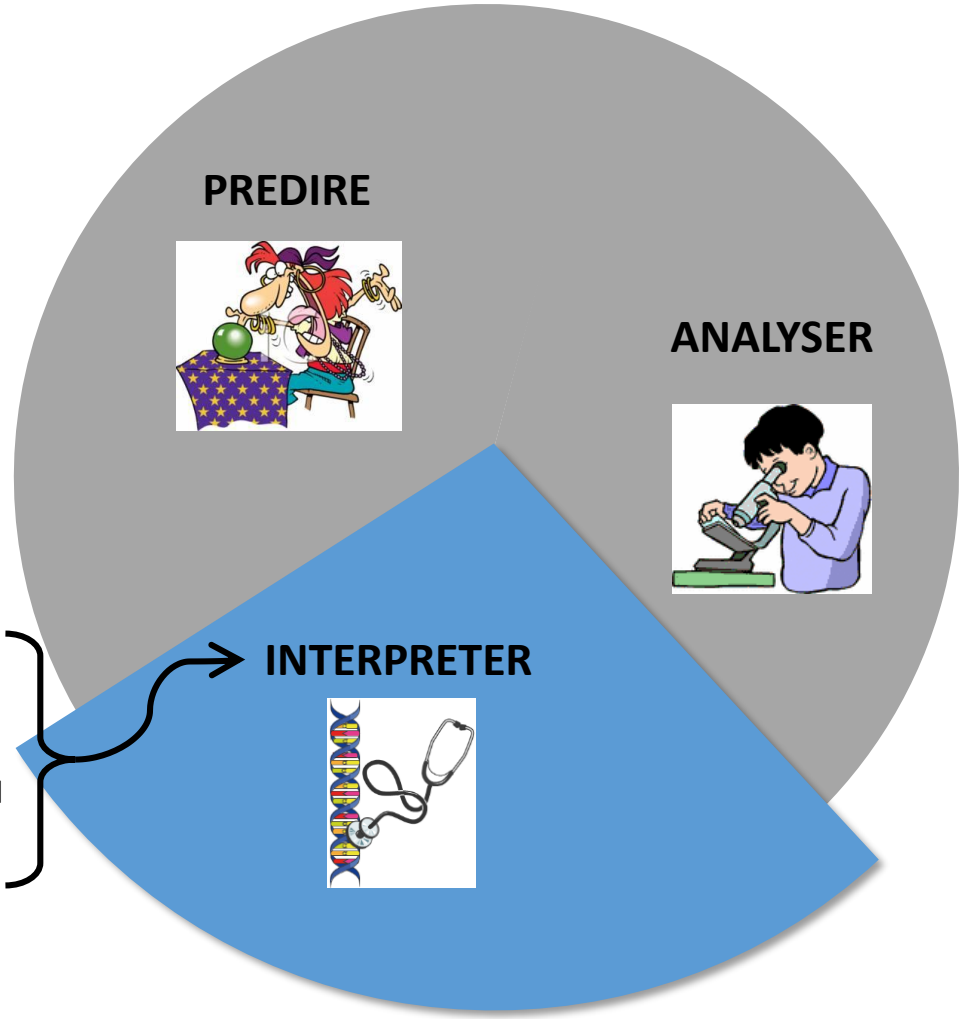
Excel file

Junction id	Gene	Read count	Relative expr.	Annotation
Chr1_123_456	*ABC1*	1 ... 5	0.1 ... 10.2	Δ3 c.23 c.230
⋮	⋮	⋮ ⋱ ⋮	⋮ ⋱ ⋮	⋮
ChrY_987_478	*ZYZ2*	10 ... 12	32.3 ... 25.4	Δ12p c.10 c.15

BRCA2 (NM_000059) > 1 % **Epissage alternatif**



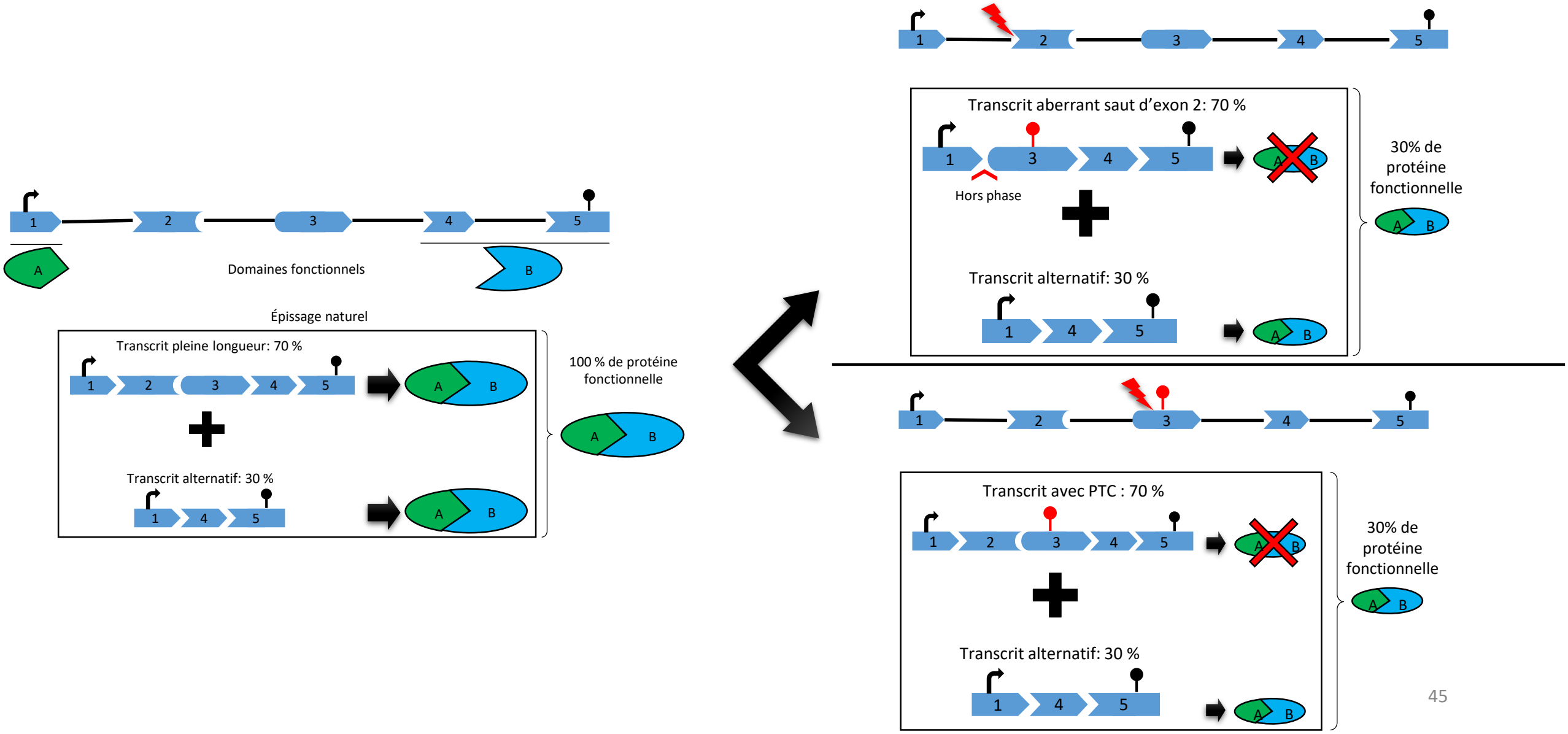
- Saut d'exon
- Site 3' alternatif
- Site 5' alternatif



Rôles de l'épissage dans l'interprétation

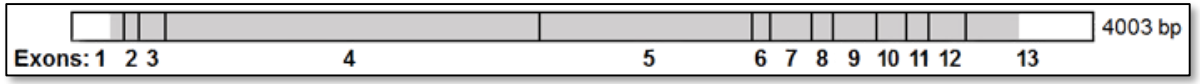
- Paradigme du gène *PALB2*, impliqué dans la prédisposition aux cancers du sein et de l'ovaire

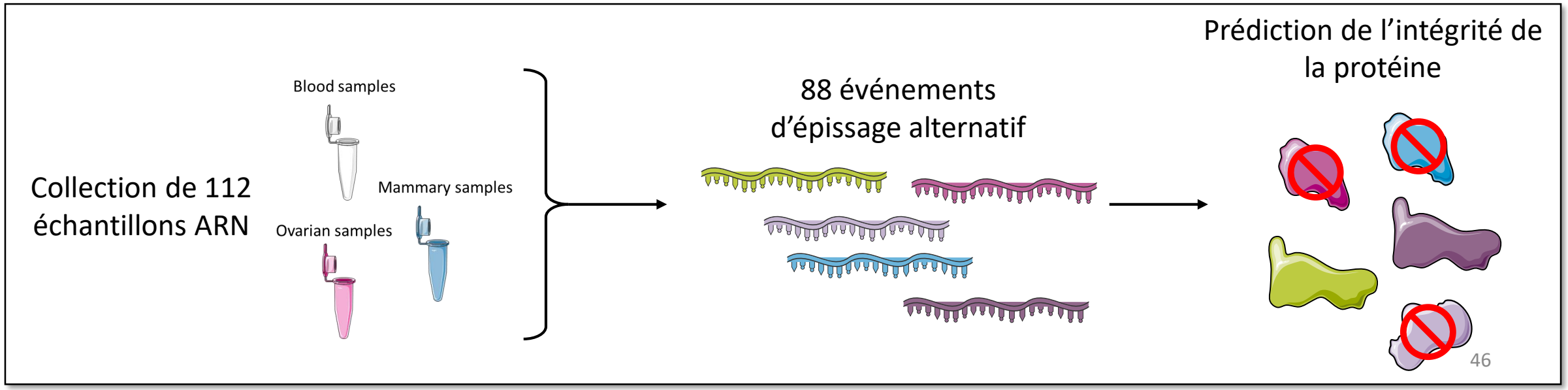
Épissage alternatif de *PALB2* et règles d'interprétation *(1/3)*



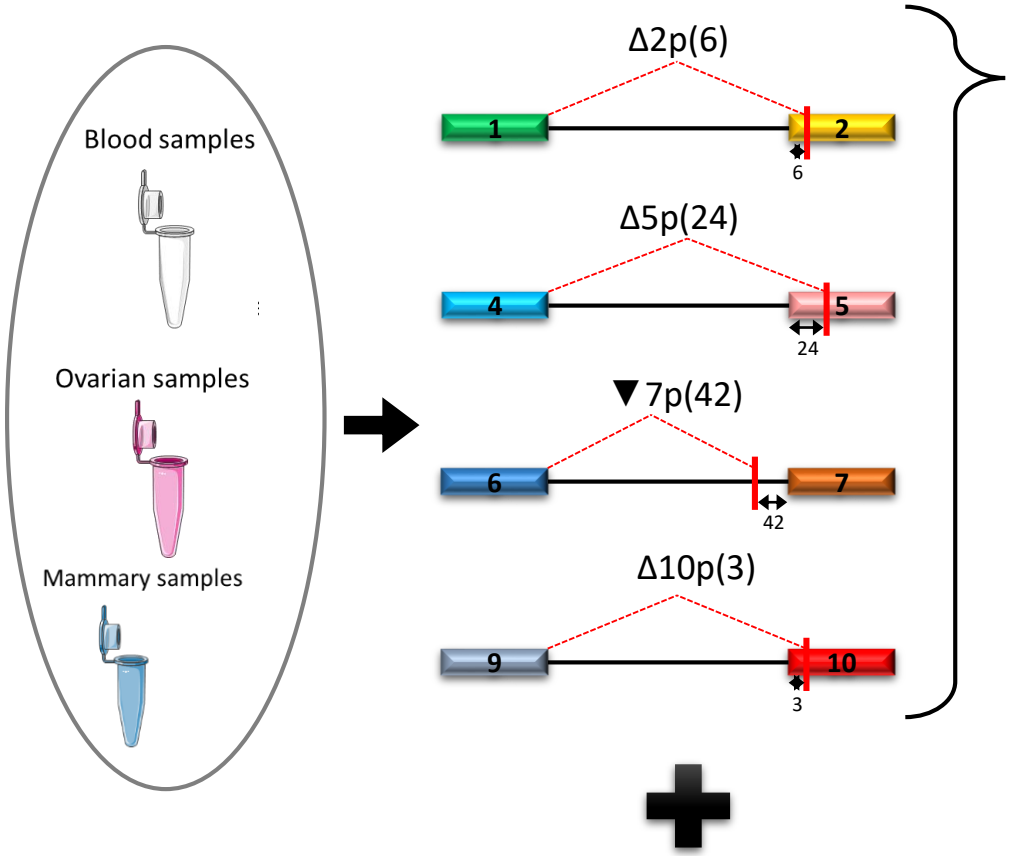
Épissage alternatif de *PALB2* et règles d'interprétation *(2/3)*

PALB2, gène de prédisposition aux cancers du sein et de l'ovaire





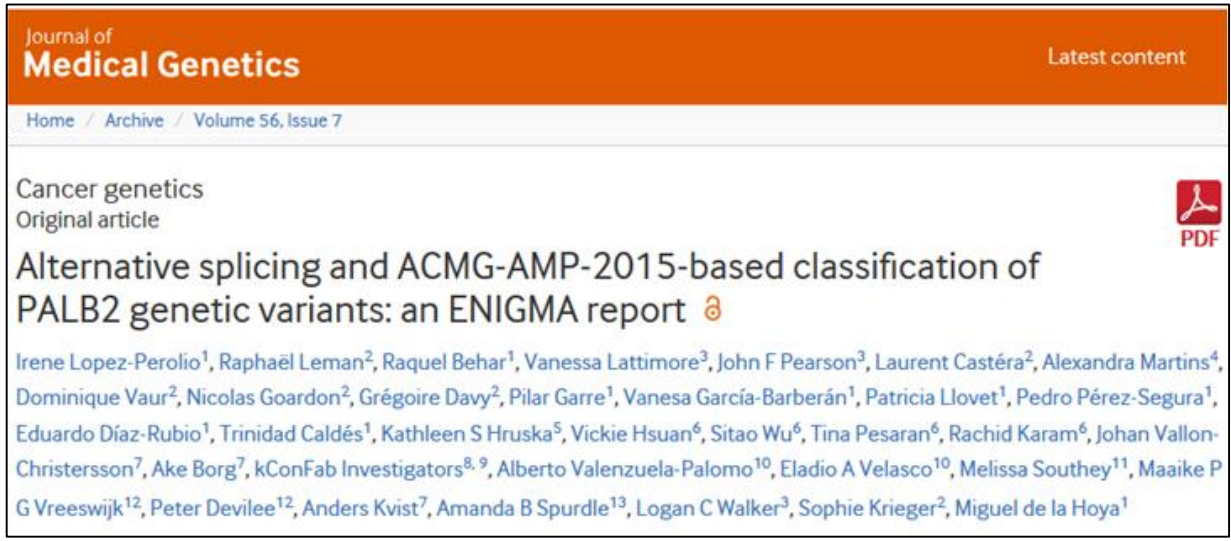
Épissage alternatif de *PALB2* et règles d'interprétation *(3/3)*



Pas d'épissage alternatif pouvant sauvegarder la protéine si variant tronquant

4 sites accepteurs pour lesquels la classe PVS n'est pas garantie

c.49-2A>T
c.1685-2A>G
c.1685-1G>C
c.2587-2A>C
c.2997-2A>C


Journal of Medical Genetics — Latest content

Home / Archive / Volume 56, Issue 7

Cancer genetics
Original article

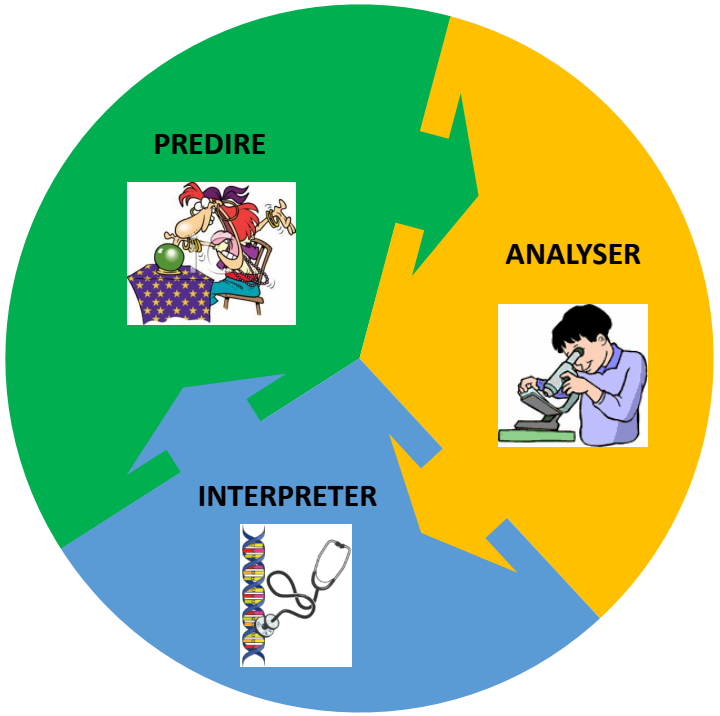
PDF

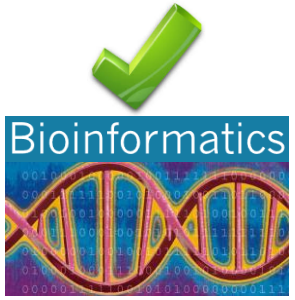
Alternative splicing and ACMG-AMP-2015-based classification of PALB2 genetic variants: an ENIGMA report

Irene Lopez-Perolio[1], Raphaël Leman[2], Raquel Behar[1], Vanessa Lattimore[3], John F Pearson[3], Laurent Castéra[2], Alexandra Martins[4], Dominique Vaur[2], Nicolas Goardon[2], Grégoire Davy[2], Pilar Garre[1], Vanesa García-Barberán[1], Patricia Llovet[1], Pedro Pérez-Segura[1], Eduardo Díaz-Rubio[1], Trinidad Caldés[1], Kathleen S Hruska[5], Vickie Hsuan[6], Sitao Wu[6], Tina Pesaran[6], Rachid Karam[6], Johan Vallon-Christersson[7], Ake Borg[7], kConFab Investigators[8, 9], Alberto Valenzuela-Palomo[10], Eladio A Velasco[10], Melissa Southey[11], Maaike P G Vreeswijk[12], Peter Devilee[12], Anders Kvist[7], Amanda B Spurdle[13], Logan C Walker[3], Sophie Krieger[2], Miguel de la Hoya[1]

Résumé des travaux de thèse

SPiP ➔ un outil pour prédire un défaut d'épissage quelle que soit la position du variant



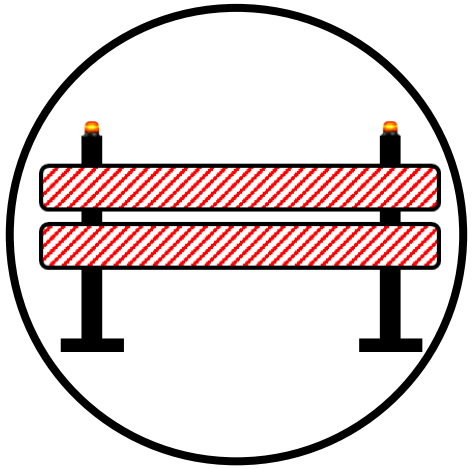


SpliceLauncher ➔ Outils prometteurs pour analyser l'épissage

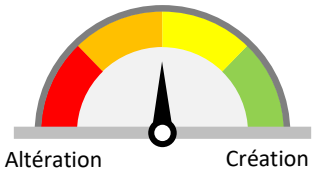
Étude de l'épissage ➔ une étape obligatoire pour l'interprétation des variants

Limites actuelles des prédictions

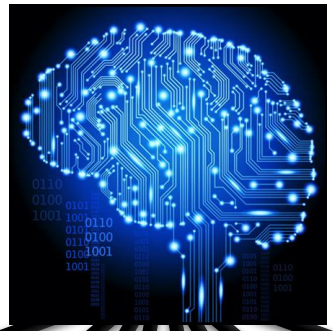
Les prédictions d'épissage ne prédisent pas la pathogénicité



Score de prédiction



Les limites à l'utilisation du *Deep Learning*



Iss-PC[1] 2017

Branchpointer[2] 2017

LaBranchoR[3] 2018

RNABPS[4] 2019

SPANR[5] 2015

SpliceAI[6] 2019

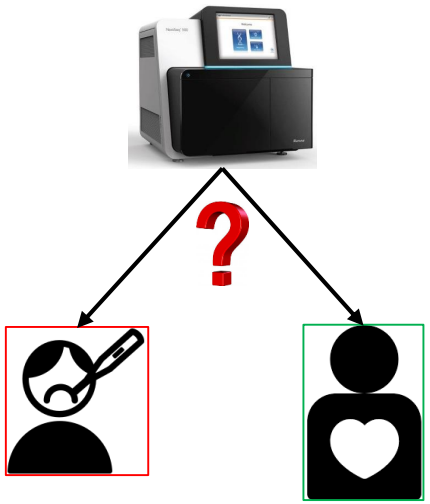
SpliceRover[7] 2018

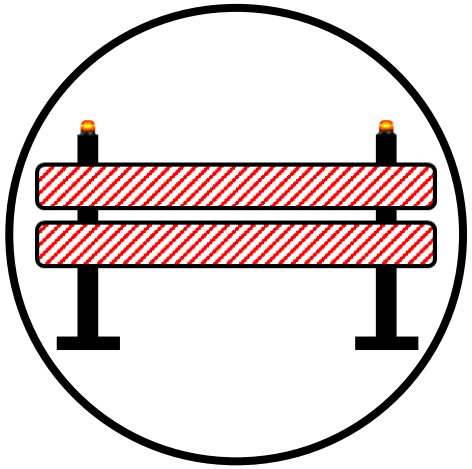
MMSplice[8] 2018

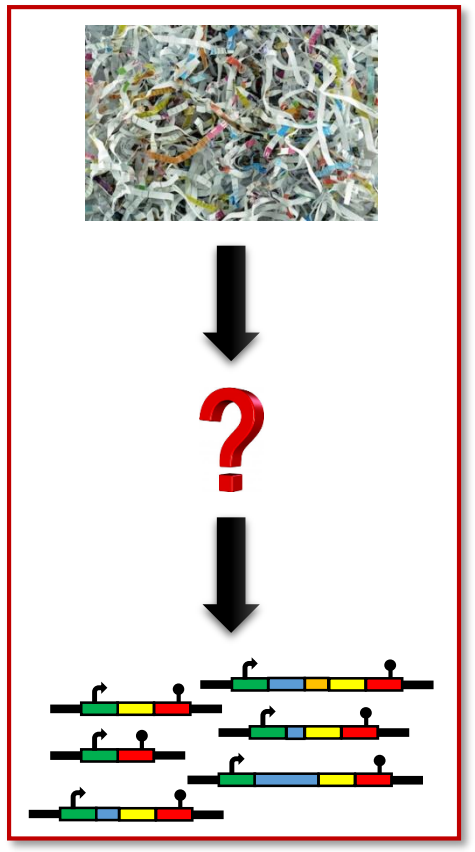
[1]Xu *et al., Sci. Rep.*, 2017
[2]Signal *et al., Bioinformatics*, 2017
[3]Paggi *et al., RNA*, 2018
[4]Nazari *et al., IEEE Access*, 2019
[5]Xiong *et al., Science*, 2015
[6]Jaganathan *et al., Cell*, 2019
[7]Zuallaert *et al., Bioinformatics*, 2018
[8]Cheng *et al., Genome Biol,* 2019

Limites actuelles des analyses par RNA-seq

Pas de recommandations pour l'utilisation du RNA-seq en diagnostic moléculaire





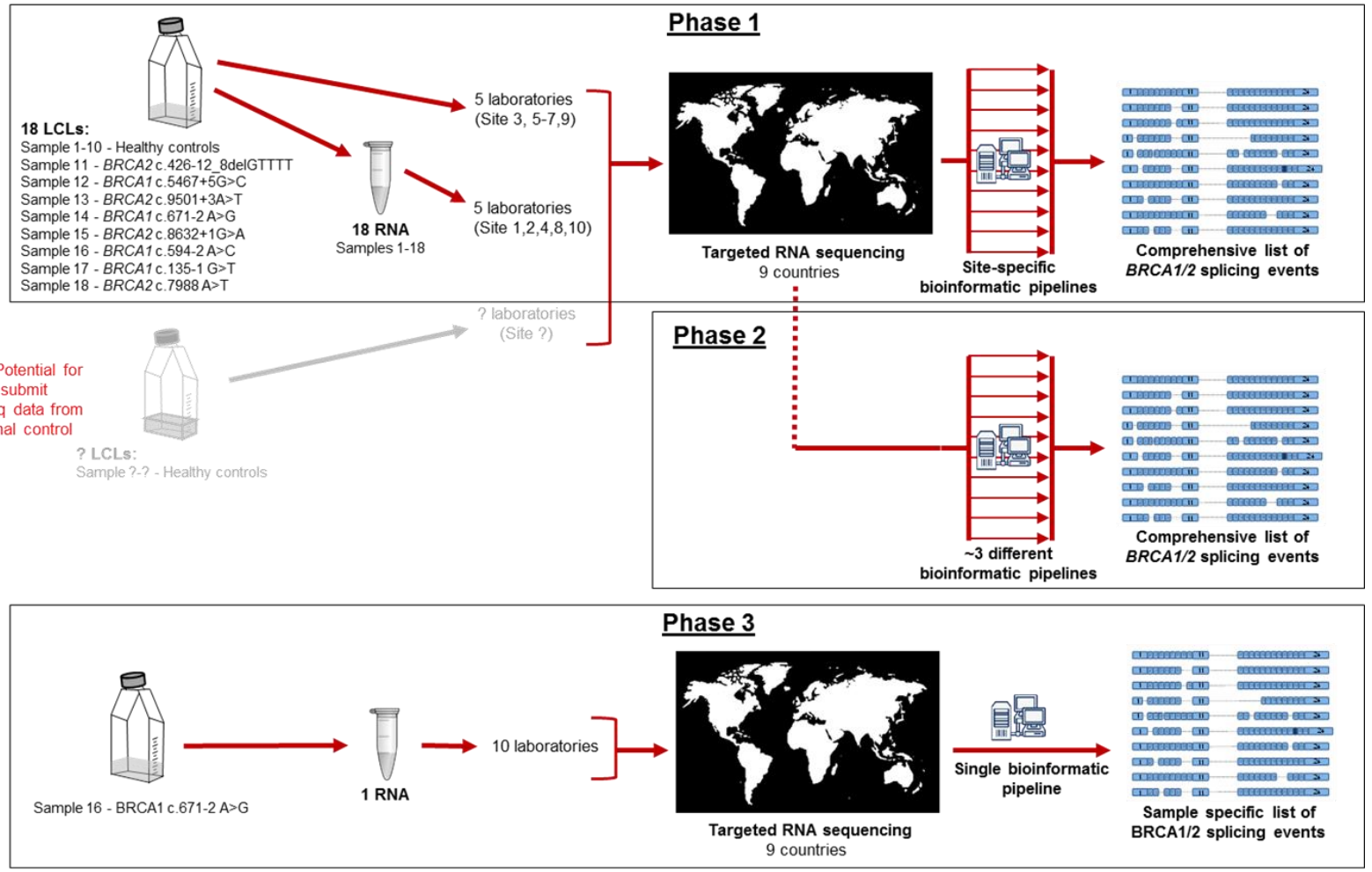


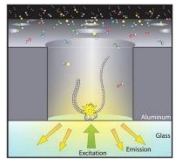
Projets en cours
– QC RNA-seq –

Établissement de recommandations pour l'analyse des défauts d'épissage par RNA-seq

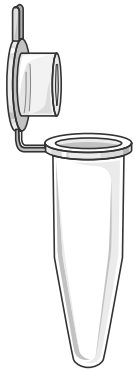
Projet mené par Logan Walker:

- 11 laboratoires
- 9 pays
- 9/11 ont fourni les données de séquençage
- Données brutes analysées par SpliceLauncher



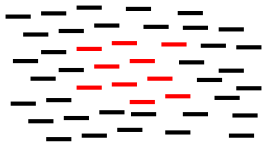


Projets en cours
– Séquençage long read pour RNA-seq ciblé –



RT
SMARTer™ PCR cDNA Kit

Clontech

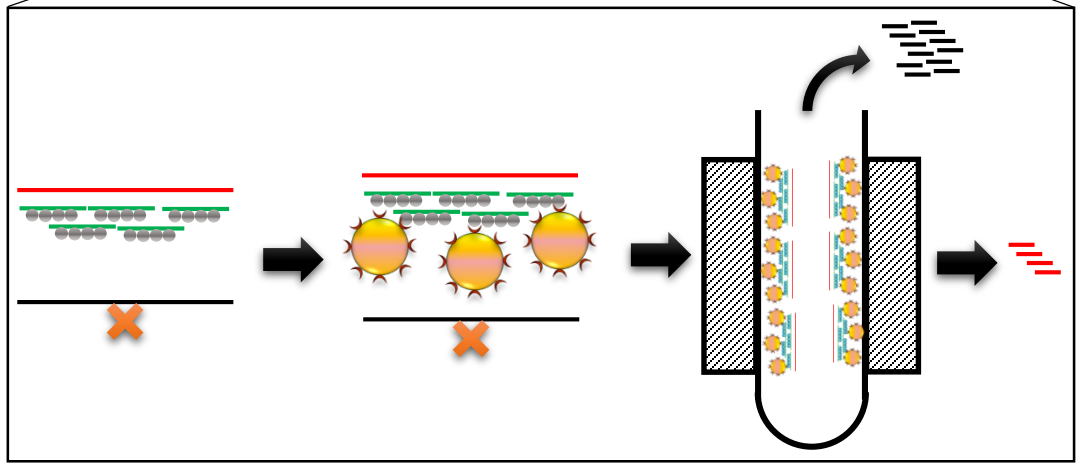


Target enrichment
SureSelect XT

Agilent

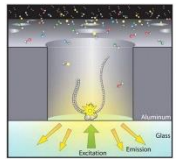
Library construction
SMRTbell

PACIFIC BIOSCIENCES



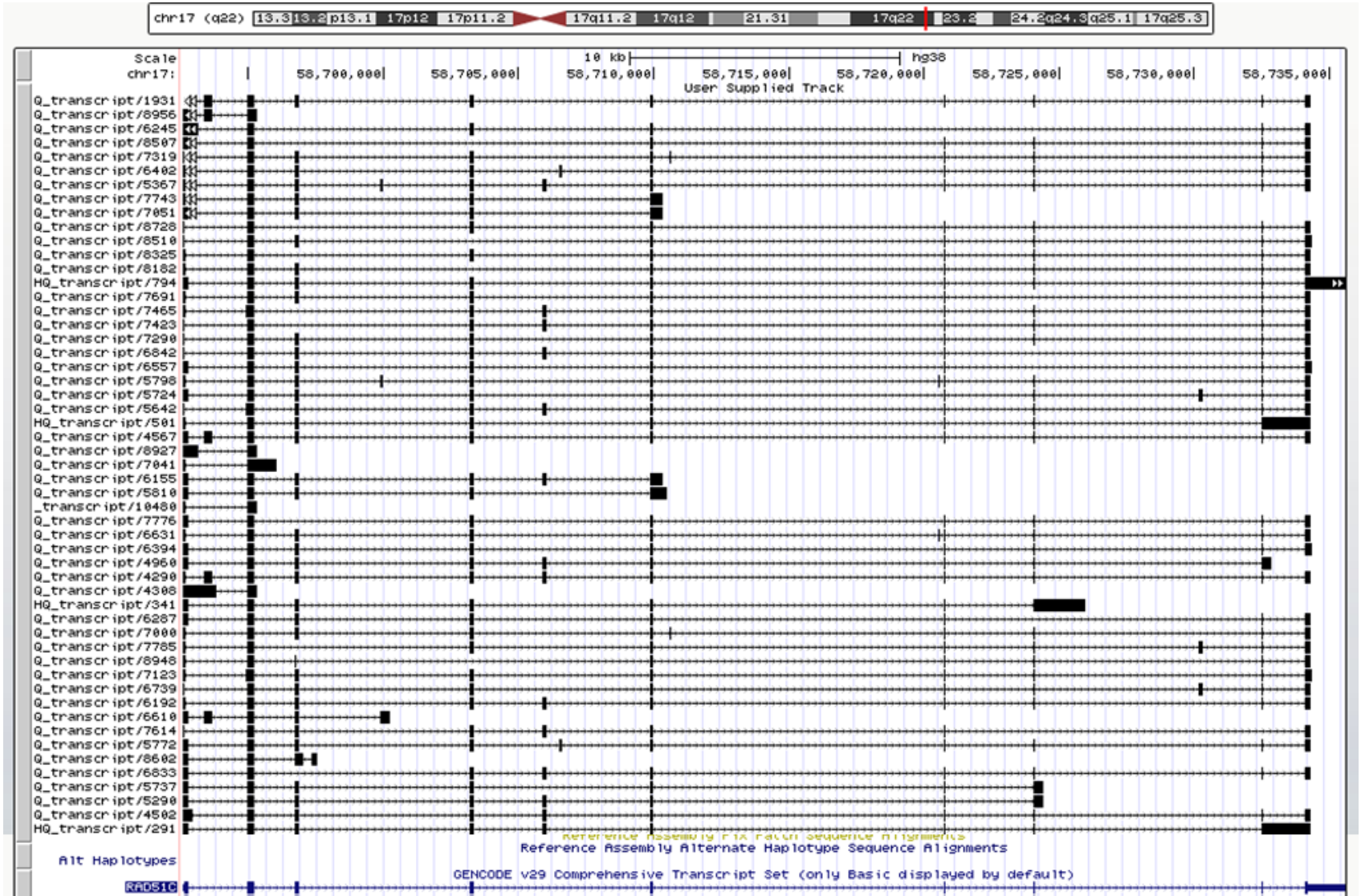
Canceropôle Nord-Ouest

20 000 €



Projets en cours
– Séquençage long read pour RNA-seq ciblé –

RAD51C 



Conclusion et perspectives

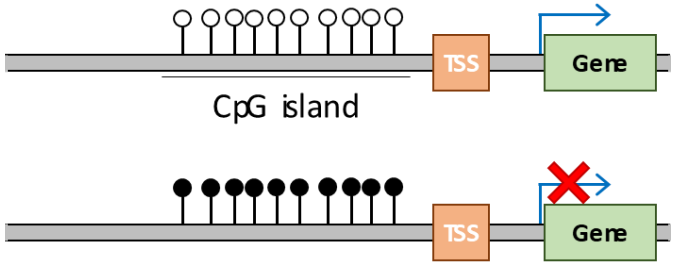
Proportion de variants utilisables pour le diagnostic moléculaire

Régions exoniques
Variants non-sens
+ faux-sens
> 65 % dans ClinVar

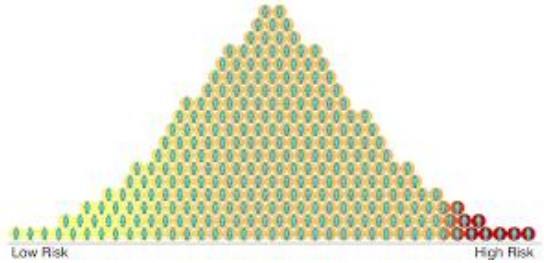
Régions « intra-géniques »
Variants splicéogéniques

Régions « inter-géniques »
????

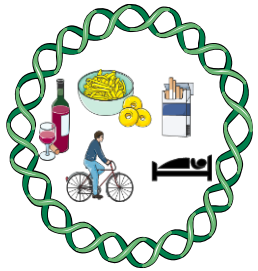
Épigénétique



Facteurs polygéniques



Gènes et environnement



Merci pour votre attention

~ The End ~

Sophie KRIEGER. . . . . Directrice de Thèse
Alexandra MARTINS. . . . . Co-directrice de Thèse

Inserm U1245

Dominique VAUR. . . . . Directeur du laboratoire de biologie, Caen
Claude HOUDAYER. . . Directeur du laboratoire de biologie, Rouen
Thierry FRÉBOURG. . . . . Directeur d'équipe, Inserm U1245

Agathe RICOU. . . . . Biologiste
Laurent CASTERA. . . . . Biologiste
Stéphanie BAERT-DESURMONT. . . . . Biologiste
Etienne MULLER. . . . . Assistant biologiste
Isabelle TOURNIER. . . . . Maitre de conférences
Angelina LEGROS. . . . . Chef d'équipe
Nicolas GOARDON. . . . . Cadre de service
Omar SOUKARIEH. . . . . Ingénieur
Pascaline GAILDRAT. . . . . Chargée de recherche
Alexandre ATKINSON. . . . . Bioinformaticien

Antoine ROUSSELIN. . . . . Bioinformaticien
Raphaël LANOS. . . . . Bioinformaticien
Céline DERAMBURE. . . . . Technicienne
Gaia CASTELAIN. . . . . Technicienne
Julie HAUCHARD. . . . . Technicienne
Audrey KILLIAN. . . . . Technicienne
Céline QUESNELL . . . . . Technicienne
Grégoire DAVY. . . . . Etudiant en thèse
Hélène TUBEUF. . . . . Etudiant en thèse
Sabine RAAD. . . . . Etudiant en thèse

Inserm UMR1078

Claude FEREC. . . . . Directeur d'équipe
Chandran KA. . . . . Ingénieur
Gérald Le GAC. . . . . Maitre de conférences
Yann FICHOU. . . . . Ingénieur
Marie-Pierre AUDREZET. . . . . Biologiste

Inserm U1016

Béatrice PARFAIT. . . . . Biologiste
Dominique VIDAUD. . . . . Biologiste
Emmanuelle GIRODON. . . . . Biologiste
Laurence PACOT. . . . . Interne

Valentin HARTER. . . . . Biostatisticien
Pierre de la GRANGE. . . . . GenoSplice
Florence RIANT. . . . . Biologiste
Fréderic LEMOINE. . . . . Bioinformaticien

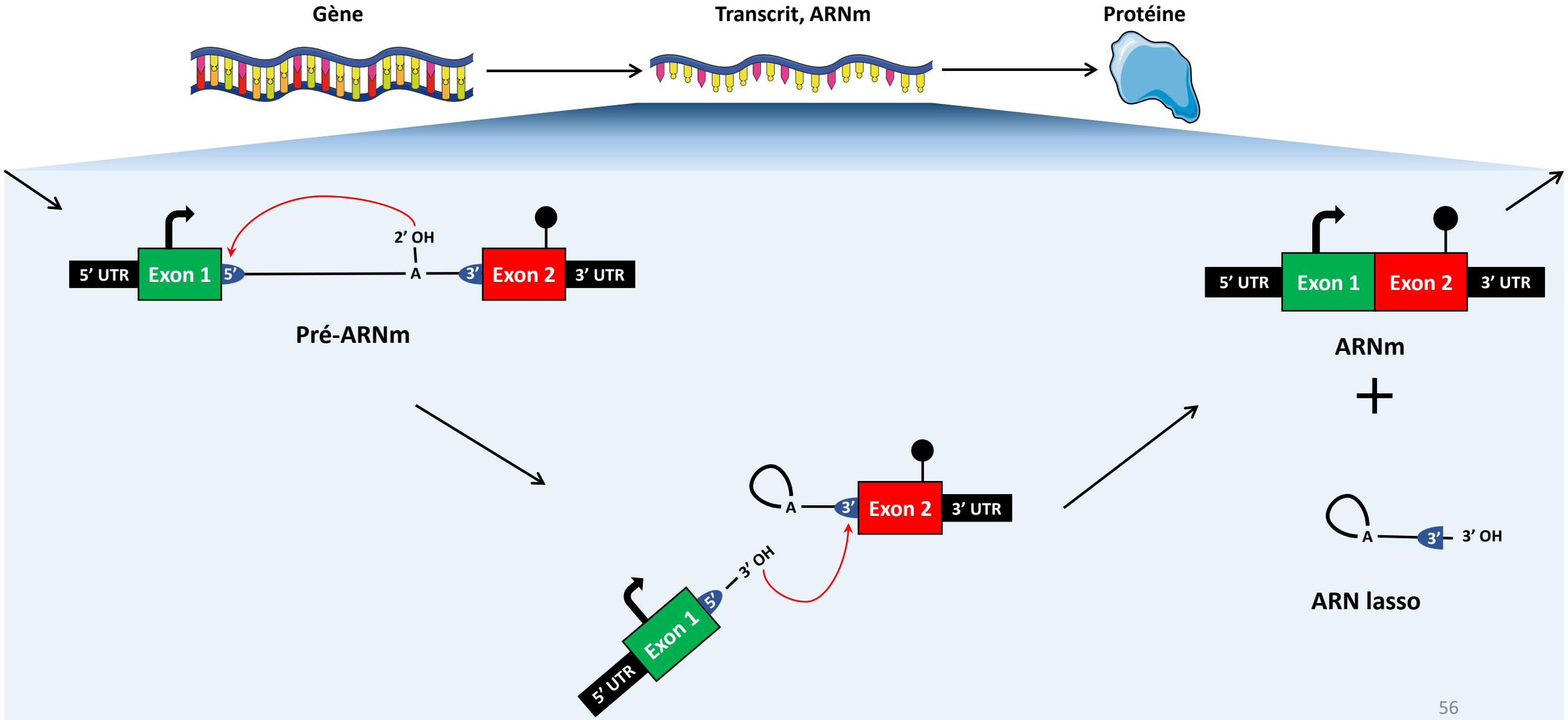
Réseau GGC

Brigitte BRESSAC-DE PAILLERETS. . . . . Biologiste
Capucine DELNATTE. . . . . Biologiste
Danielle MULLER. . . . . Biologiste
Etienne ROULEAU. . . . . Biologiste
Florence COULET. . . . . Biologiste
Virginie CAUX-MONCOUTIER. . . . . Ingénieur
Violaine BOURDON. . . . . Biologiste
Sylvie MAZOYER. . . . . Biologiste
Sandrine M. CAPUTO. . . . . Chargée de recherche
Nicolas SEVENET. . . . . Biologiste
Françoise BONNET-DORION. . . . . Ingénieur
Mélanie LÉONE. . . . . Biologiste
Marine GUILLAUD-BATAILLE. . . . . Ingénieur
Françoise REVILLON. . . . . Biologiste
Nadia BOUTRY-KRYZA. . . . . Biologiste
Joanna SOKOLOWSKA. . . . . Biologiste
Inès SCHULTZ. . . . . Biologiste

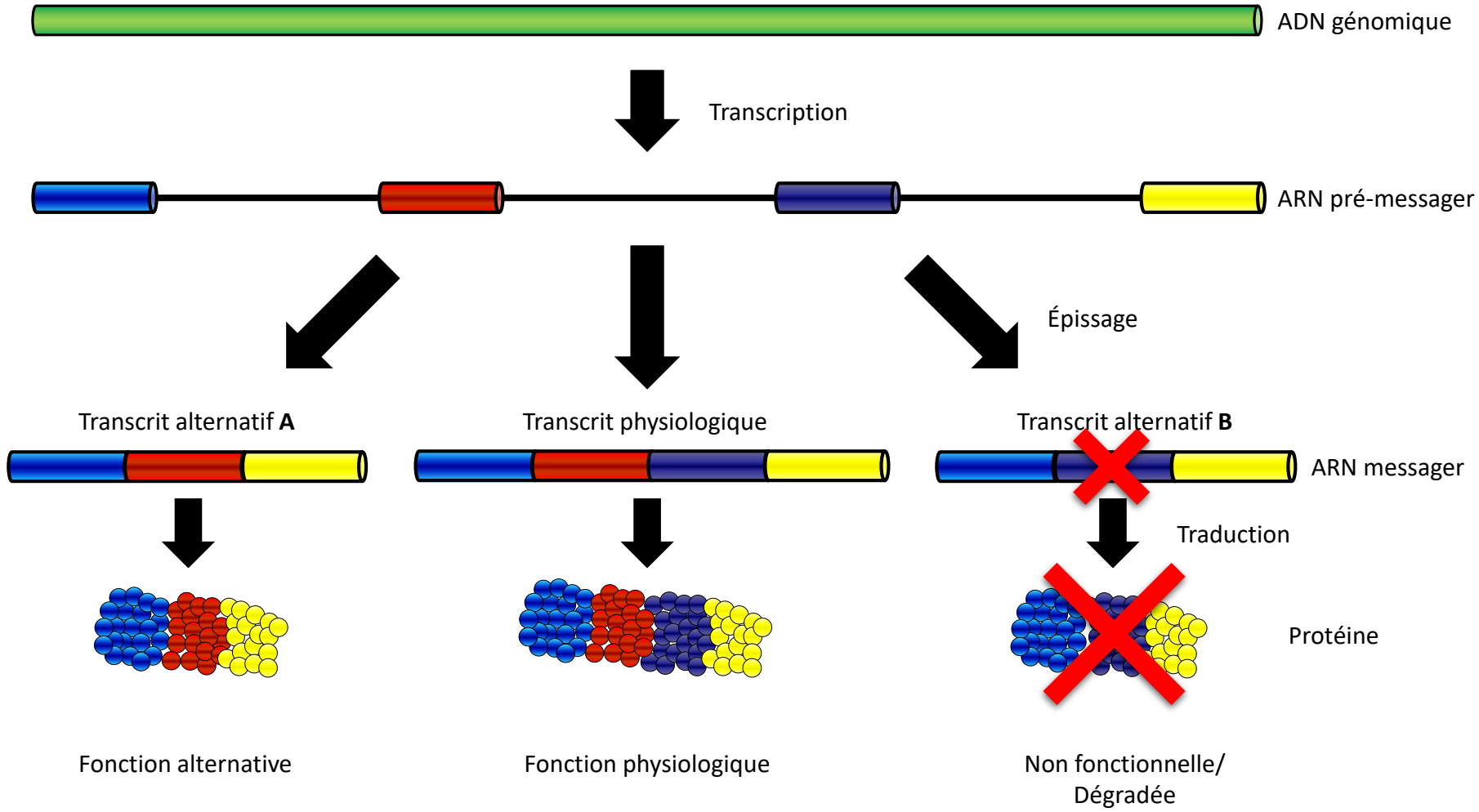
ENIGMA

Amanda B. SPURDLE. . . . . Professor
Anders KVIST. . . . . Researcher
Barbara WAPPENSCHMIDT. . . . . Fachhumangenetikerin
Irene LOPEZ-PEROLIO . . . . . Research Associate
Logan C WALKER. . . . . Associate Professor
Maaike P G VREESWIJK. . . . . Researcher
Maria ROSSING. . . . . Researcher
Melissa SOUTHEY. . . . . Professor
Michael T. PARSONS. . . . . Research assistant
Miguel de la HOYA. . . . . Biologiste
Rachid KARAM. . . . . Ambry genetics
Vanessa LATTIMORE. . . . . Research Fellow

Mécanisme d'épissage



L'épissage alternatif

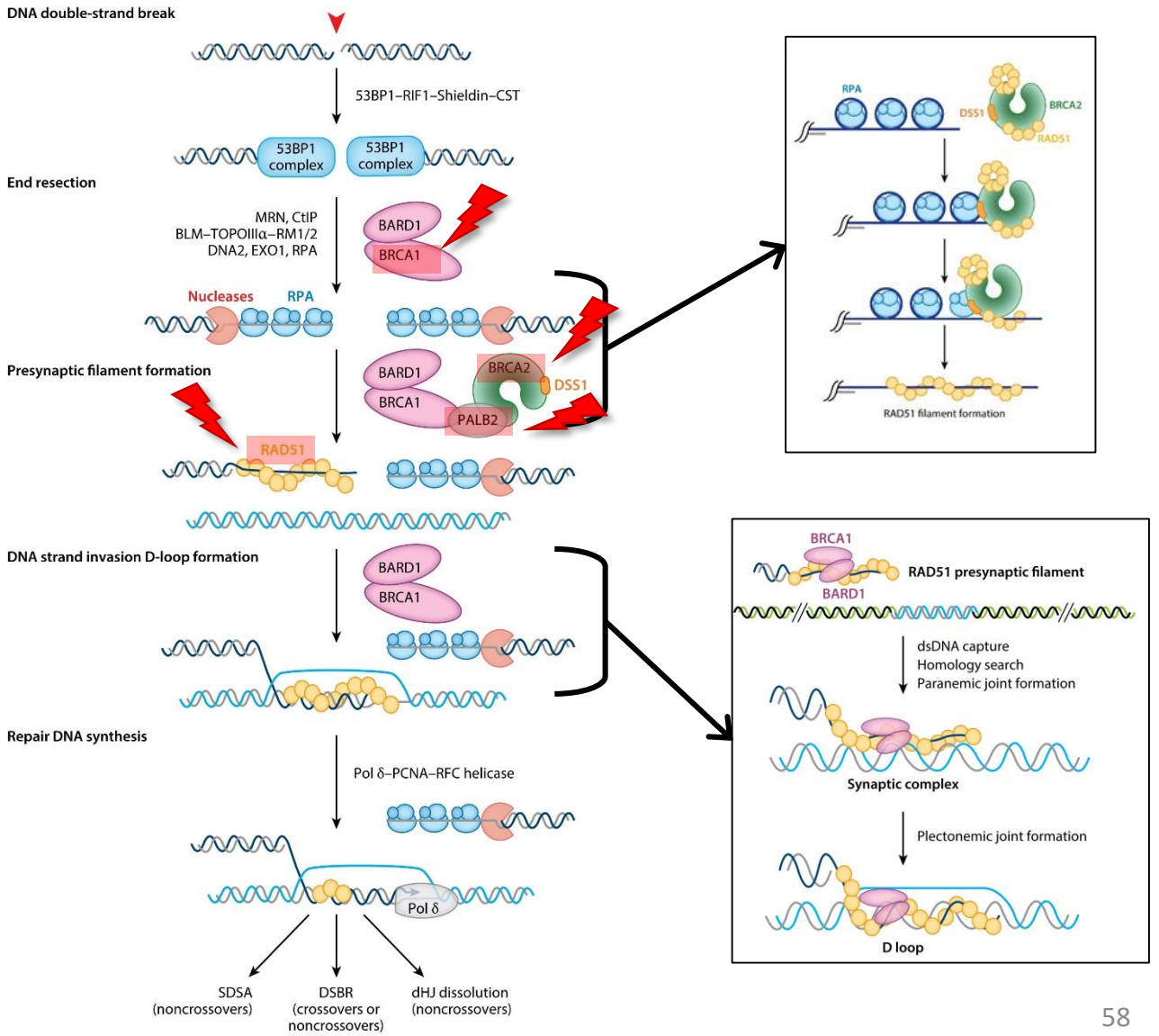


Gènes impliqués dans le syndrome HBOC

5 genes majeurs de predisposition :

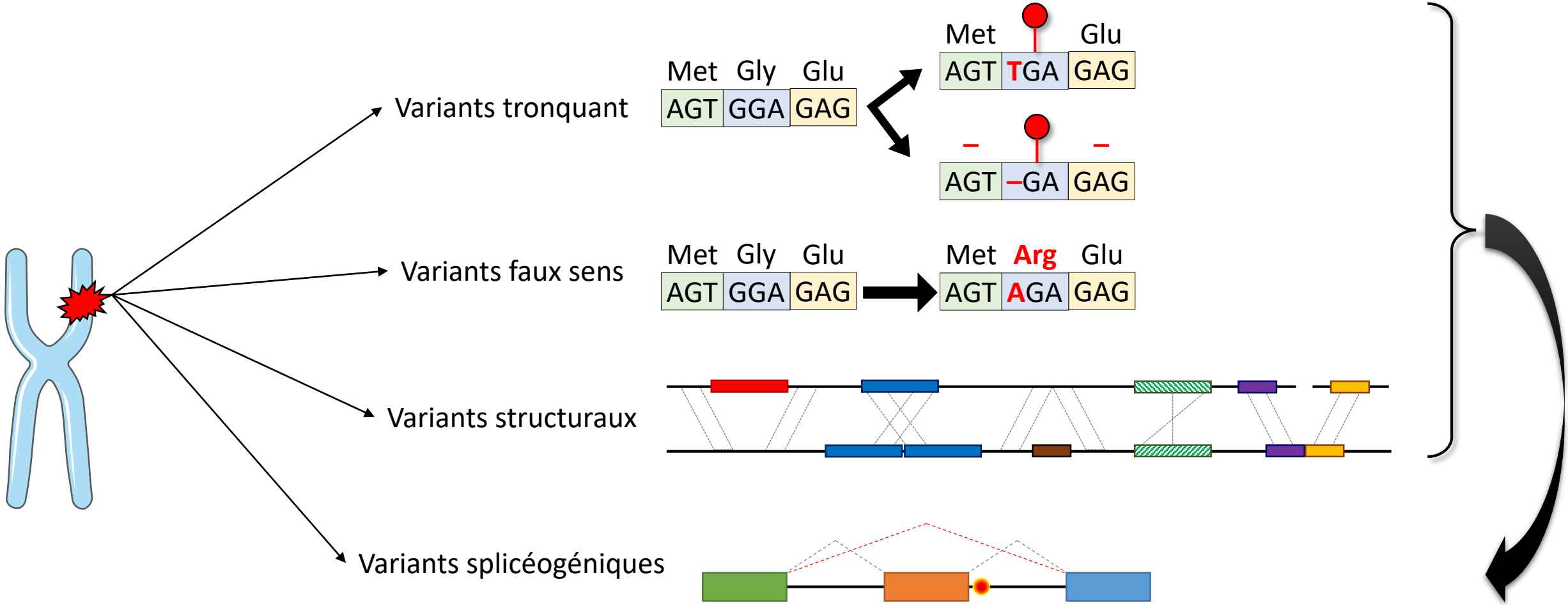
- *BRCA1* (1994)
- *BRCA2* (1995)
- *PALB2* (2007)
- *RAD51C* (2010)
- *RAD51D* (2011)

Recombinaison homologue



Zhao *et al., Annu. Rev. Biochem.*, 2019

Variation génétique et consequence sur la protéine



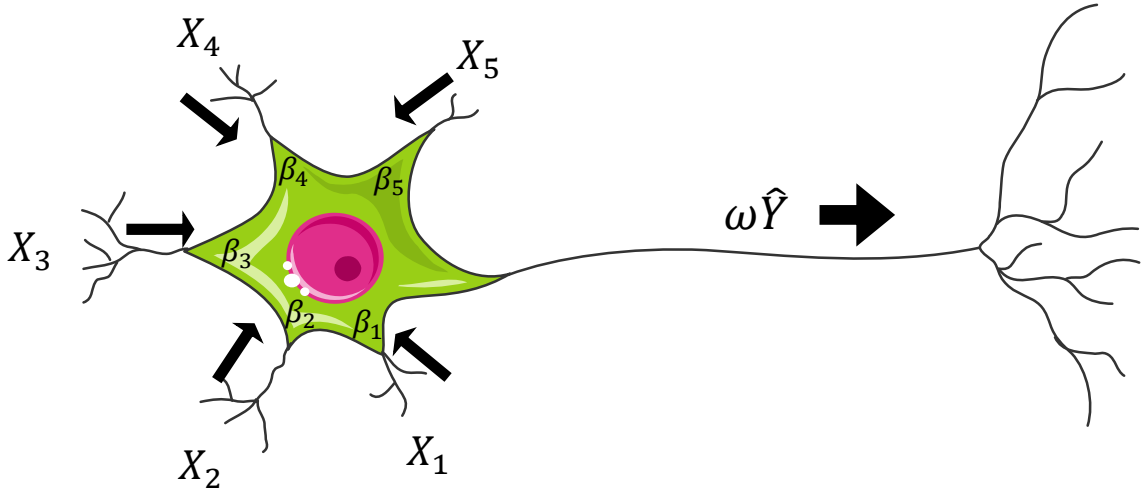
Intelligence artificielle

1950's | 1960's | 1970's | 1980's | 1990's | 2000's | 2010's

Machine learning

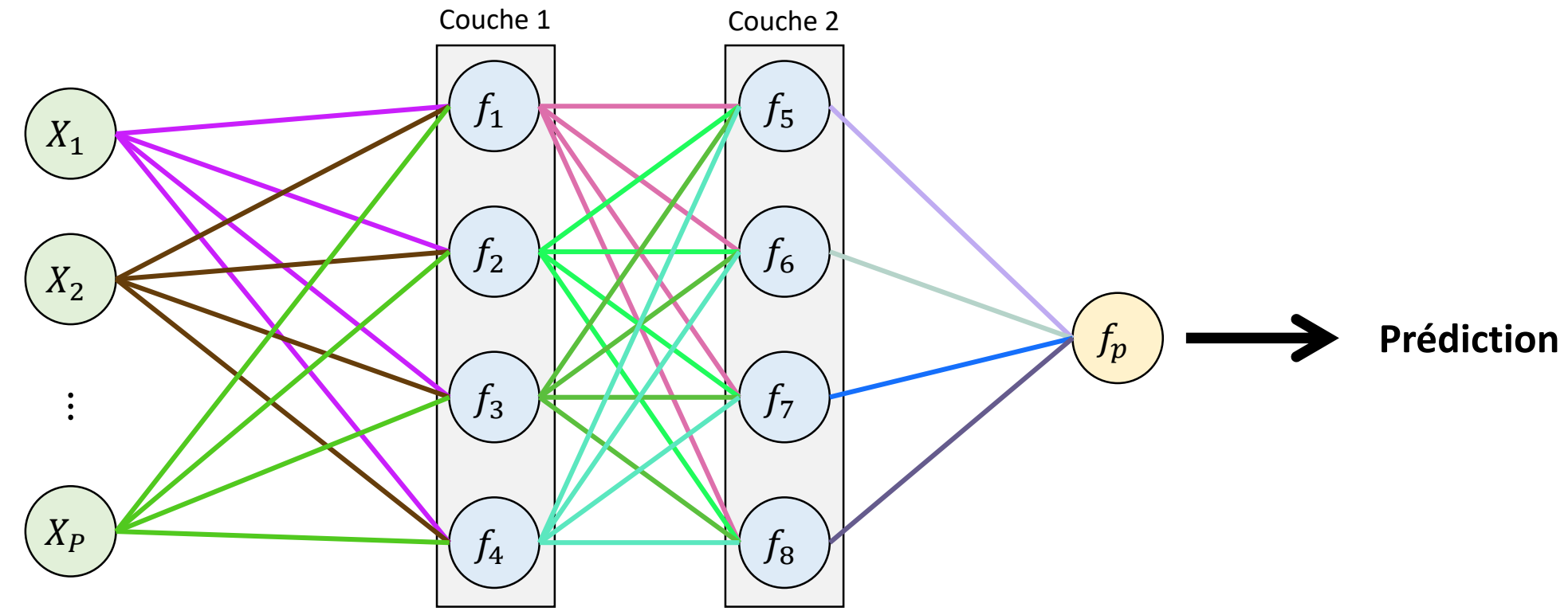
Deep learning

$$\hat{Y} = \beta_1 \times X_1 + \beta_2 \times X_2 + \cdots + \beta_p \times X_p$$



Machine learning

Les réseaux de neurones



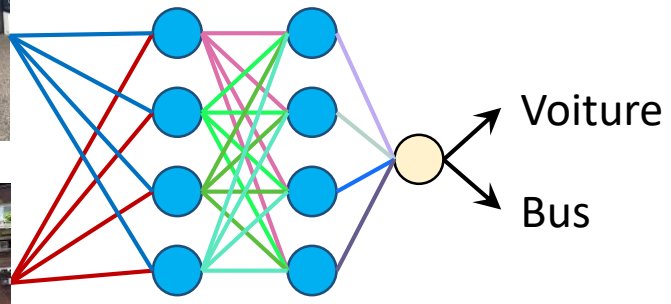
Entrées **Neurones** **Neurone final**

L'apport du *deep learning*

Machine learning

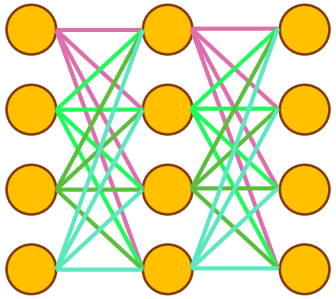


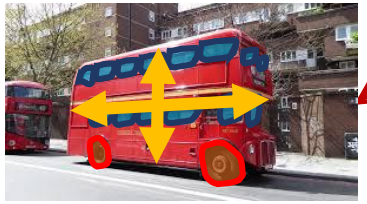
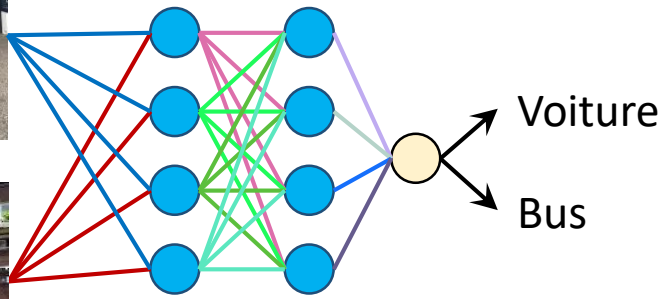


Deep learning

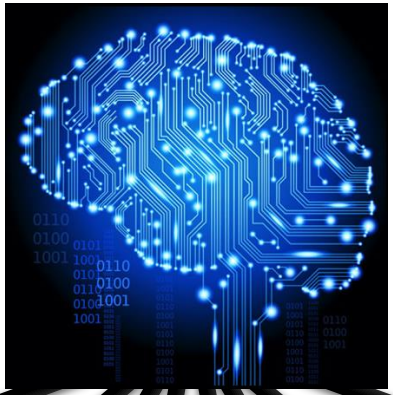




Deep learning et prédiction épissage

Méthode en pleine essor



Iss-PC[1] 2017

Branchpointer[2] 2017

LaBranchoR[3] 2018

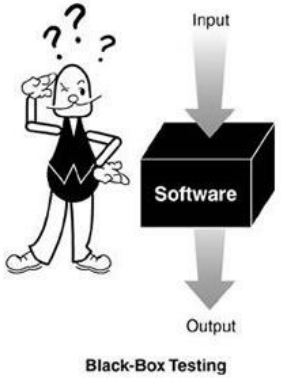
RNABPS[4] 2019

SPANR[5] 2015

SpliceAI[6] 2019

SpliceRover[7] 2018

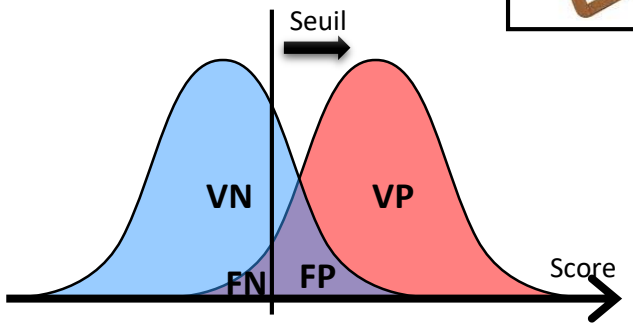
MMSplice[8] 2018



[1]Xu *et al., Sci. Rep.,* 2017
[2]Signal *et al., Bioinformatics*, 2017
[3]Paggi *et al., RNA*, 2018
[4]Nazari *et al., IEEE Access*, 2019
[5]Xiong *et al., Science*, 2015
[6]Jaganathan *et al., Cell*, 2019
[7]Zuallaert *et al., Bioinformatics*, 2018
[8]Cheng *et al., Genome Biol,* 2019

Évaluation des outils de prédiction

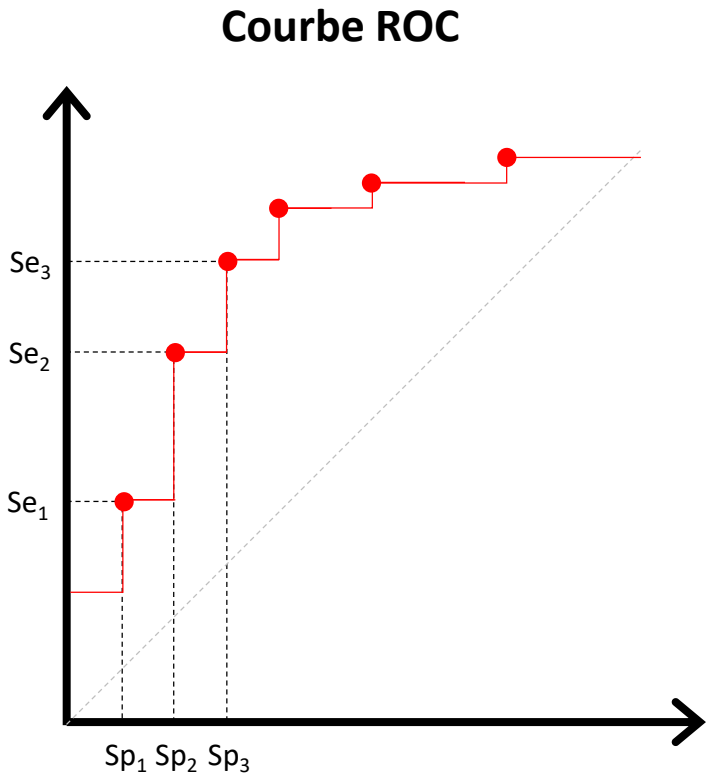
Collection de variants avec études ARN *in vitro*

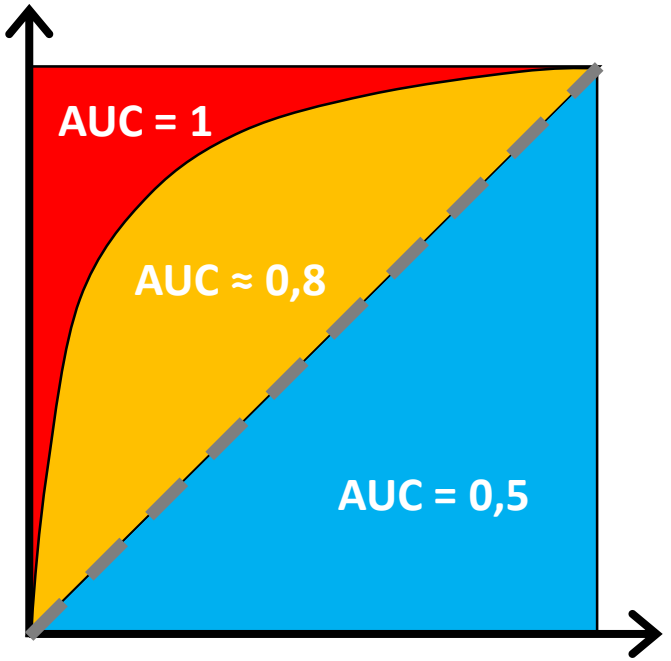


Seuil 1 Se_1/Sp_1		Observation	
		OUI	NON
Prédiction	OUI	VP	FP
	NON	FN	VN

Seuil 2 Se_2/Sp_2		Observation	
		OUI	NON
Prédiction	OUI	VP	FP
	NON	FN	VN

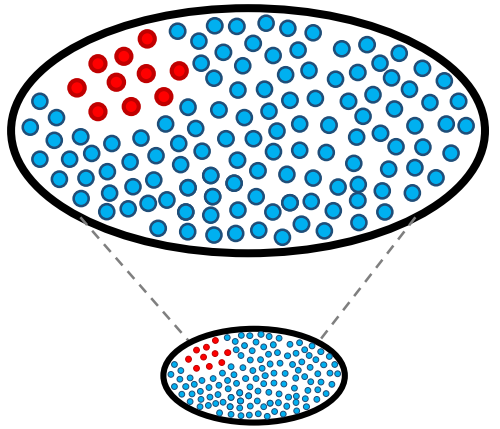
⋮

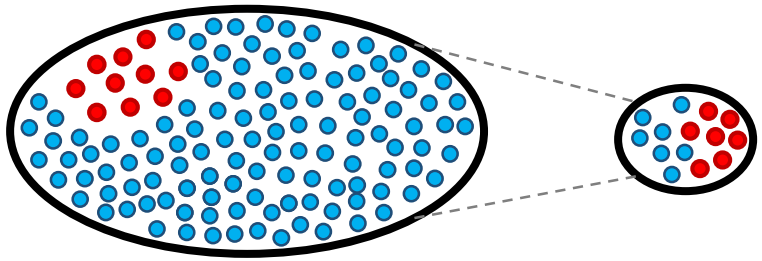




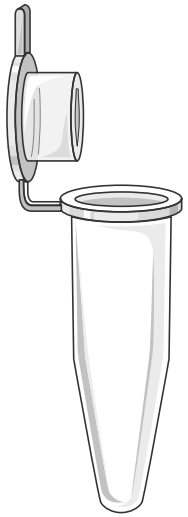
Évaluation des outils de prédiction

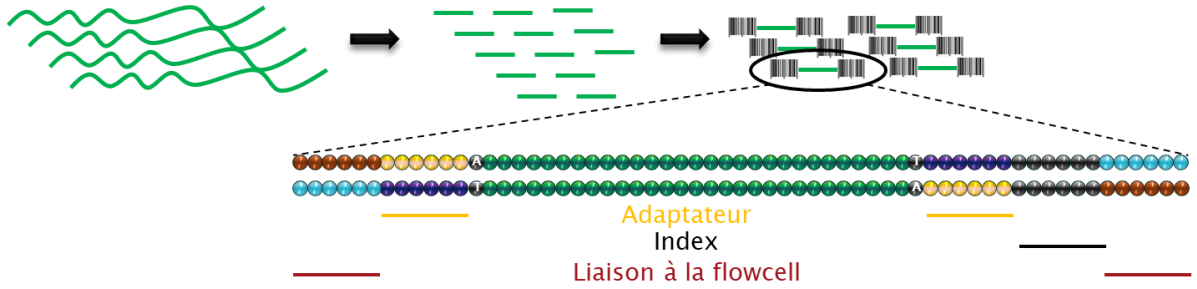
	Population totale	Observation : Positive	Observation : Négative	
Prédiction	Positive	**Vrai positif, VP**	**Faux positif, FP**	**Valeur prédictive positive (VPP), Précision** $= \frac{VP}{VP + FP}$
Prédiction	Négative	**Faux négatif, FN**	**Vrai négatif, VN**	**Valeur prédictive négative (VPN)** $= \frac{VN}{VN + FN}$
		***True positive rate* (TPR), *Recall*, Sensibilité, Puissance** $= \frac{VP}{VP + FN}$	**Spécificité, Sélectivité, *True negative rate* (TNR)** $= \frac{VN}{VN + FP}$	**Exactitude (*accuracy*)** $= \frac{VP + VN}{VP + FP + VN + FN}$

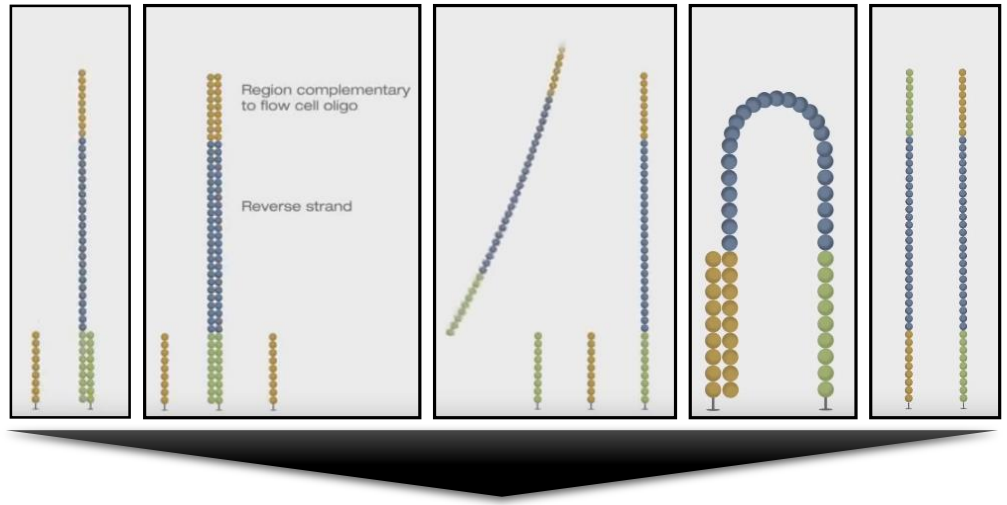


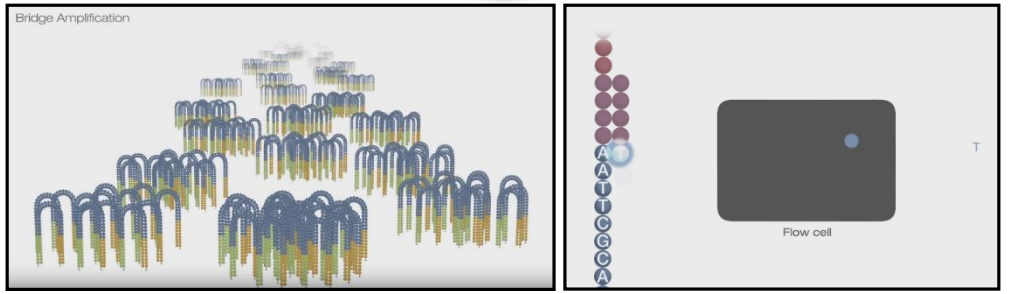


Principe RNA-seq, ex : Illumina®





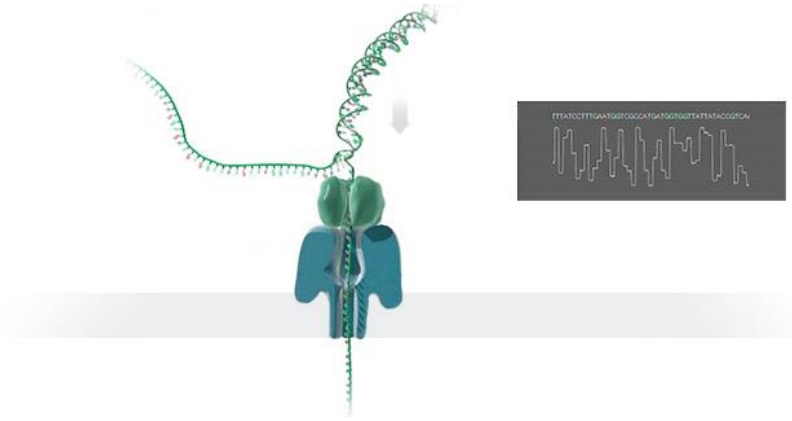


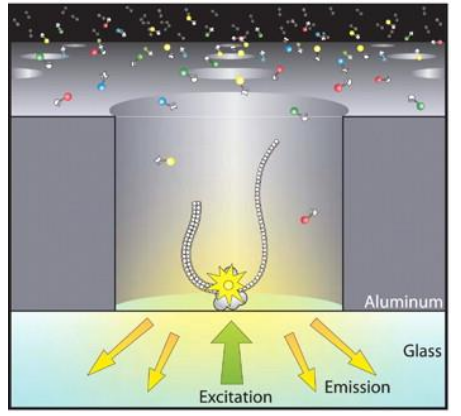


RNA-seq long read

Séquençage molécule unique

Transcrit de 1 à 100 kb

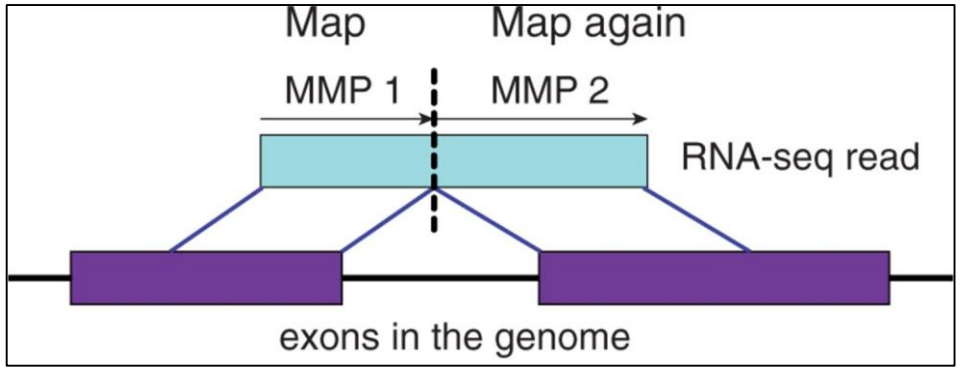




RNA-seq et analyses bioinformatiques

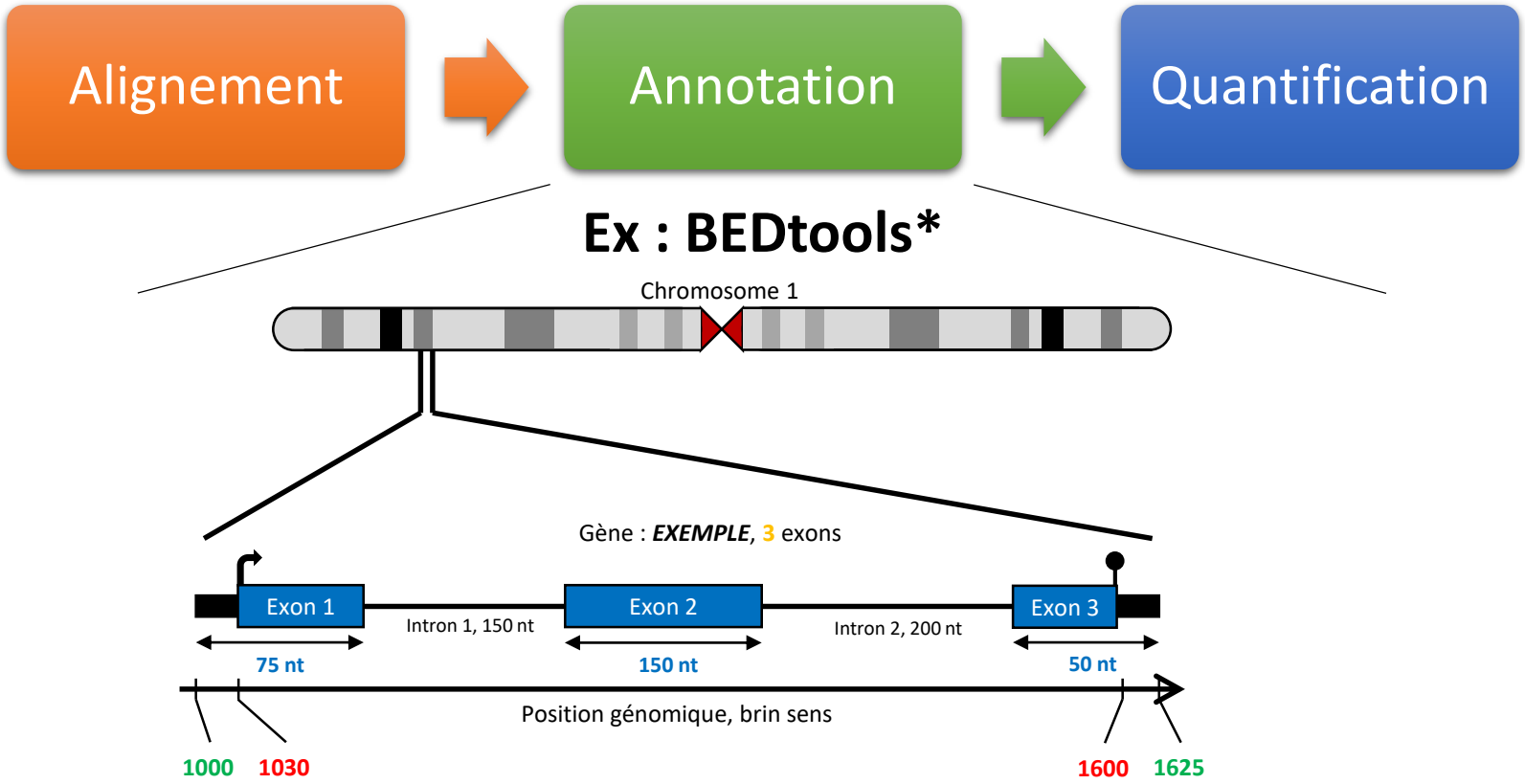
Alignement → Annotation → Quantification

Ex : STAR*



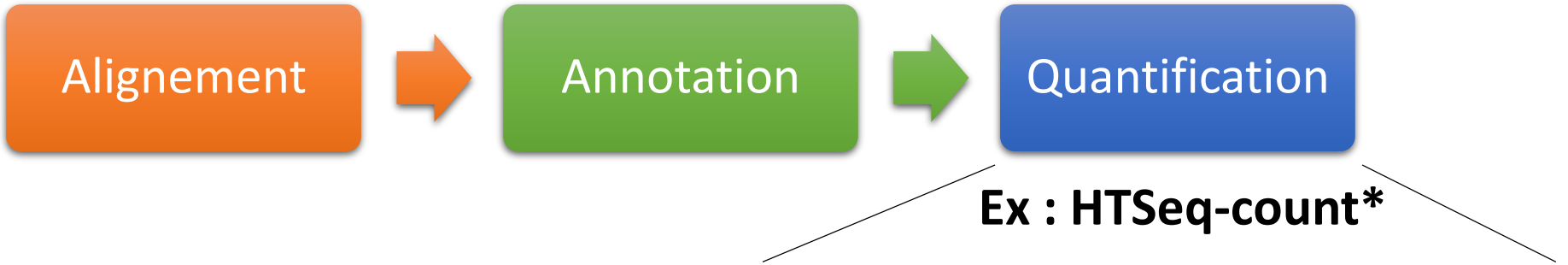
*Dobin *et al., Bioinformatics*, 2013

RNA-seq et analyses bioinformatiques



*Quinlan *et al.*, *Bioinformatics*, 2010

RNA-seq et analyses bioinformatiques

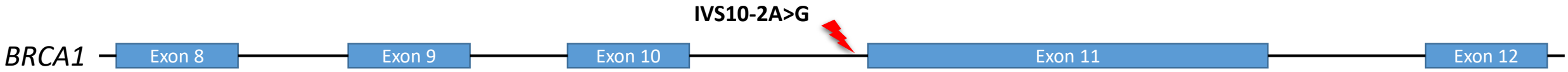


	Ech_1	Ech_2	$\cdots$	Ech_j
Iso_1	C_{11}	C_{12}	$\cdots$	C_{1j}
Iso_2	C_{21}	C_{22}	$\cdots$	C_{2j}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Iso_i	C_{i1}	C_{i2}	$\cdots$	C_{ij}

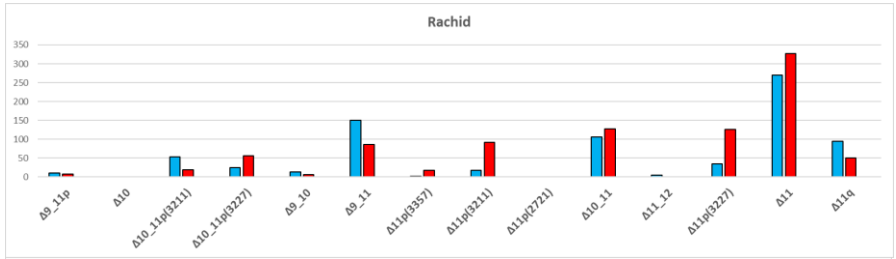
TO BE CONTINUED... SpliceLauncher

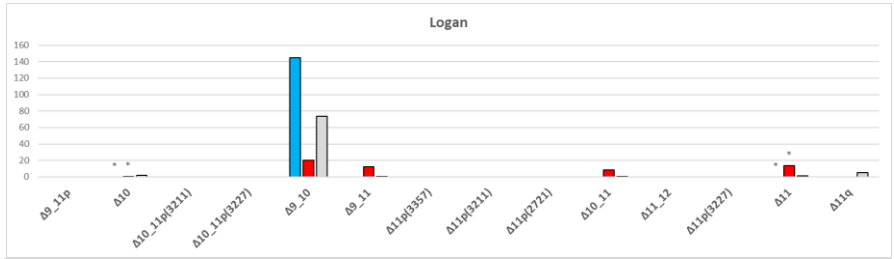
*Anders *et al.*, *Bioinformatics*, 2015

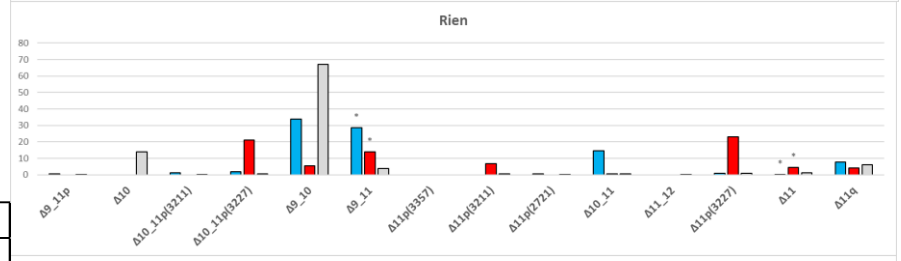
QC RNA-seq: towards guidelines for RNA-seq analyses *(2/2)*

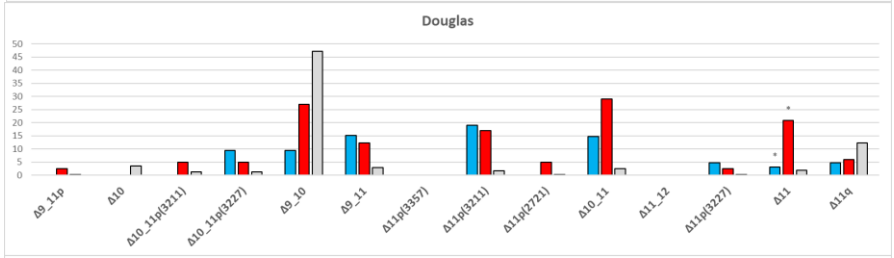


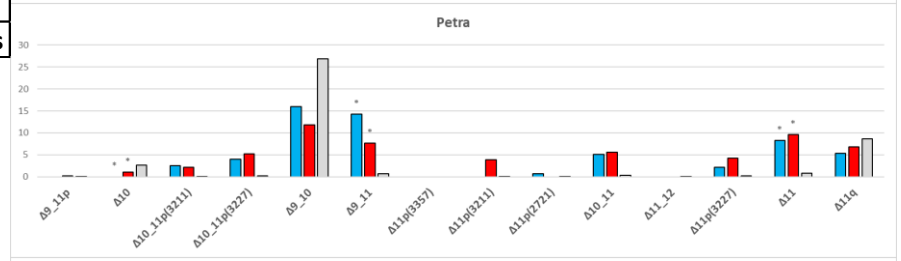
	IVS10-2 untreated
	IVS10-2 treated
	Mean accross samples

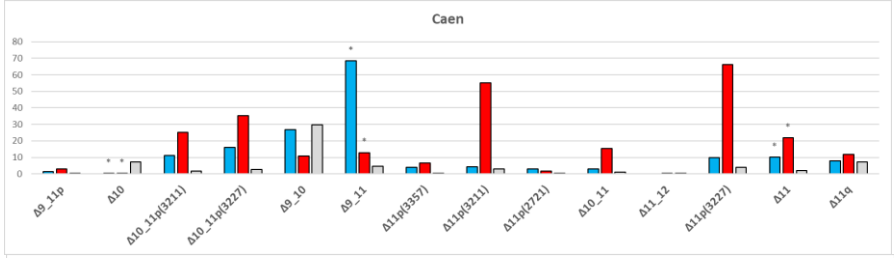


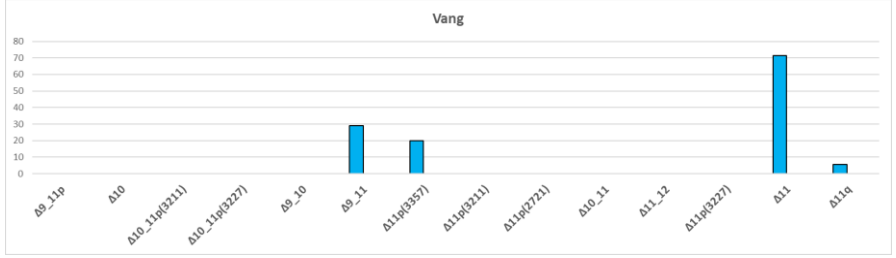


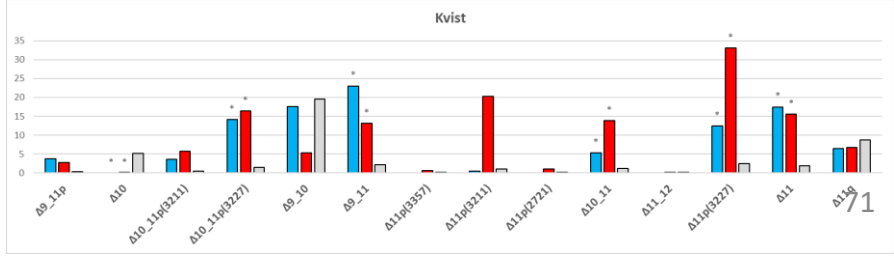




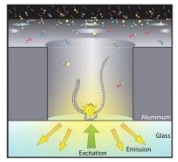






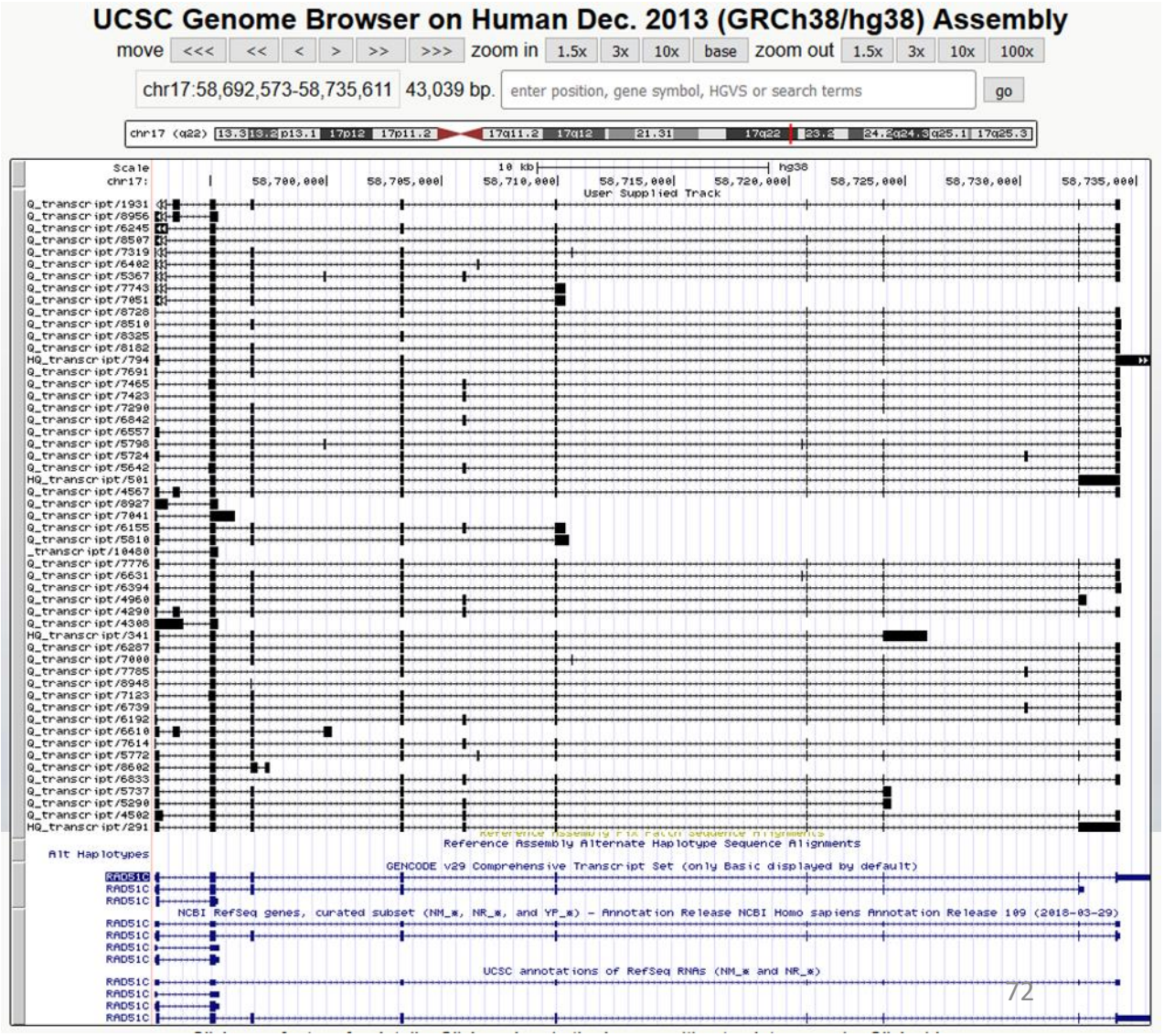


Long read sequencing for targeted RNA-seq *(3/3)*



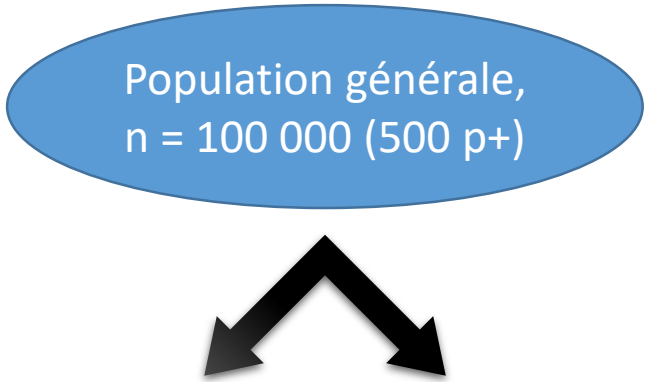
First results

Read count	*ACTB*	*BRCA1*	*BRCA2*	*PALB2*	*RAD51C*	*RAD51D*
Without capture	2 986	5	4	3	3	0
With capture	26	277	203	139	1107	14



Le piège des faux positifs

ACMG rules :
Se : 20 %
Sp : 99 %



Population **avec** histoire familiale

N = 900	P+ (450)	P- (450)
T+	90	~5
T-	360	445

Tests positifs < 100

Population générale

N = 100 000	P+ (500)	P- (99 500)
T+	100	995
T-	400	98 505

Tests positifs > 1 000

Fausse conclusion

Faire un test génétique sur une population générale permet de trouver dix fois plus de prédisposition qu'un test selon les recommandations actuelles